Gradient-Guided Annealing for Domain Generalization

Supplementary Material

The following materials are provided in this supplementary file:

- An extended literature review discussion, helpful for navigating the Domain Generalization literature under the scope of computer vision.
- A computational analysis regarding the application of GGA.
- Detailed results for each dataset domain and algorithm, presented in Table 2 of the main text, along with extra experiments on the ColoredMNIST and RotatedMNIST datasets.

A. Extended Literature Review

Domain Generalization (DG) [46] is arguably one of the most difficult and fundamental problems of Machine Learning (ML) today. Unsurprisingly, a vast number of researchers have poured effort into advancing the field, where findings have been applied to various areas, such as Natural Language Processing [22], Reinforcement Learning [29], Healthcare and Medicine [3, 15], Time-Series forecasting [16], Fault Diagnosis [56] and, of course, Computer Vision [46]. Even though not covering the entire field of DG, this section aims to present a taxonomy of the general DG methodologies developed in CV, for producing robust models that can generalize to previously unseen data, and attempts to assist potential readers navigate the past literature, while also categorizing our proposed method among its predecessors. Domain Generalization methods can be categorized into three major groups, depending on their operation during the process of model training, namely: (a) Data Manipulation, (b) Representation Learning and, (c) Learning Algorithm. Furthermore, as mentioned in the main text, DG algorithms can either leverage domain labels during training (multi-source), or completely disregard the knowledge of existing domain shifts in their training data and handle them as a single distribution (single-source).

Data Manipulation. As its name suggests, methods included in this group focus on either perturbing existing samples (*data augmentation*) or creating novel ones (*data generation*), in order to regularize the training of machine learning models, avoid overfitting and improve their generalizability. The basic idea in data manipulation methodologies is to simulate domain shift by creating diverse data samples, which can in turn mimic the entirety of distributions present in the input space. Regarding data augmentation, most popular techniques include traditional image transformations, such as random flip, rotation and color distortion. Even though these augmentations can be randomly applied during training, without needing domain labels, it has been shown that their selection significantly affects model performance. For example, the authors of [44] define novel augmentation rules that push the perturbed images to diverge as much as possible from the original ones. Additionally, image augmentations prove effective towards overcoming domain shifts in medical image classification [35, 54], where transformations can replicate shifts caused by the use of different devices. On the other hand, multiple data augmentation methods were also inspired by adversarial attacks and use adversarial gradients to distort the input images [37, 45], or use randomly initialized convolutional networks for transforming samples [13]. These techniques act as regularizers during model training, allowing them to learn generalizable image representations. The generation of novel data domains is also a well researched area in the data manipulation group. In addition to using domain gradients for synthesizing novel domains [39], several methods took advantage of style transfer [20] and either map the styles of images to that existing source domains [8] or create novel styles [52]. On a similar note, mixing the styles of training images by conventional methods [50, 58] or with the generative models [47] also proves beneficial.

Representation Learning. This group of methods is arguably the most prominent in DG and has been the central focus of ML [6]. Following the formulation in the main text, given a labeling function h that maps input observations $\boldsymbol{x}$ to their labels y, we can decompose it into $h = f \circ g$, where g is a parametric function that learns representations of x and f is the classifier function. The goal of representation learning can be summarized as follows:

$$\min_{f,g} \mathbb{E}_{x,y} \ell(f(g(\boldsymbol{x}; \boldsymbol{\theta})), y) + \lambda \ell_{\text{reg}} \tag{1}$$

where ℓ the loss function to be minimized and ℓ_{reg} a regularizer. Methods included in this group, focus on learning a robust and generalizable representation learning function g. The algorithms included in this group can be further categorized into three sub-groups. *Feature disentanglement* [53] methods intend to extract disentangled feature vectors from samples, where each dimension can be linked to a subset of data generating factors. The main idea is to produce a model that extracts a representation that can be further decomposed into domain-specific, domain-invariant, and class-specific features. To that end, the authors of [36] present CSD, which jointly learns a domain-invariant and domain-specific component in the final embedding and enables the extraction of disentangled representations, whereas the authors of [11] propose learning domain specific masks during training to improve model ro-

bustness. Generative models have also been proposed in the disentangled representation learning literature for DG, with variational autoencoders (VAEs) and GANs [12] being utilized for learning distinct latent subspaces for class- and domain-specific features [23]. Another promising category of methods aiming to produce disentangled representations is that of Causality-Inspired algorithms. In causal representation learning, a domain shift can be thought of as an intervention, subsequentially leading the development of models that aim to uncover the true causal data generating factors. Naturally, the prediction of a model should not be affected by interventions on spuriously correlated but irrelevant features, such as the background, color or style of the image. Under this causal consideration, the authors of [32] propose a structural causal model in order to model within-class variations and leverage the fact that inputs across domains should have the same representation, given that they derive from the same object. Similar to disentangled representations, there have been proposed methods in the literature that focus on completely disregarding domain-relevant from the final feature vectors, deriving solely domain-invariant representations. Based on the initial findings of [5], numerous works have presented algorithms that aim to minimize the representation differences across multiple source domains within a specific feature space, ensuring they become domain invariant, ultimately enabling the trained model to effectively generalize to previously unseen domains. In one of the most notable previous works in this category, Arjovsky et al. [1] enforce the optimal classifier on top of the representation space to be the same across domains and simultaneously minimize the loss across distributions. The above idea of Invariant Risk Minimization (IRM) has been extended to several other works. For example, the authors of [26] propose minimizing the variance of source-domain risks, by minimizing their extrapolated risk, while the authors of [55] propose adapting to domain shift and producing invariant representations. Finally, an alternative route towards learning generalized representations is via regularization strategies. The most representative group of methods in this category is *Gradient-Based operations*, which utilize gradient information during model training. In [21], the authors propose learning robust representations by discarding the most dominant gradients in each training iteration under the assumption that they are correlated with domain-specific features present in the source data. Another popular strategy is to seek for flat minima [10, 18] in the loss landscape of neural networks during training, assuming that models that converge to flat minima exhibit increased generalization capabilities [48, 59]. What's more, Shi et al. [40] hypothesized that gradients among domains should match and proposed an approximation of a loss inducing the maximization of the gradient inner product during training. Our method (GGA) can be categorized in this group of gradient operations, as it considers the similarity of domain gradients in the early iterations of model training and seeks for sets in the parameter space with increased gradient alignment, before continuing the optimization procedure.

Learning Algorithm. In addition to manipulating the input space or feature extractor, DG methods were also researched under the scope of alternative ML learning paradigms, such as *ensemble*, *meta*, *domain-adversarial*, *self-supervised* and *reinforcement* learning. In this section we present the most exemplary works in each category. *Ensemble-Learning* in DG initially combined several copies of the same network, each of which is trained on a specific domain [14, 57]. Alternatively, instead of using several networks, [51] proposed sharing shallow layers among CNNs. During inference, the final prediction is produced by either simple [57] or weighted averaging [49]. In *Meta-Learning* for DG, Li et al. [29] propose MLDG and split the source domains into meta-train and meta-test splits to mimic the effects of domain shift during training. Similarly, [2] proposes learning a meta regularizer for the classifier, while MAML [17] was proposed for improved parameter initialization. Another approach is that of *Adversarial Learning* (AL). In the context of DG, the aim of adversarial learning is to train a classifier to distinguish between source domains [33] and ultimately learn domain-agnostic features from the samples that can be generalized to unseen data [31]. Other learning paradigms such as *Self-Supervised* learning have also been explored in DG, which leverages unlabeled data samples to derive generalized representations. Notably, the authors of [9] introduce a self-supervised jigsaw-solving puzzle task to push the model to learn robust representations. Furthermore, contrastive learning has also been shown to improve model performance. Specifically, SelfReg [24] utilizes self-supervised contrastive losses to bring latent representations of same-class samples closer. Similarly, the authors of [4] introduce a contrastive loss for representations extracted from intermediate layers of the network. Finally, *Reinforcement learning* has also been applied in the context of DG. Indicatively, previous works have explored randomizing the environments of an RL agent for transferring them to real-world scenarios [28, 42], whereas [27] researches the combination of RL with contrastive learning.

B. Computational Analysis

B.1. Experiment Infrastructure

Each and every experiment is conducted on a cluster containing 4×40GB NVIDIA A100 GPU cards, split into 8 20GB virtual MIG devices and 1×24GB NVIDIA RTX A5000 GPU card, via a SLURM workload manager.

B.2. Complexity Analysis

Each GGA training iteration includes computing model gradients $S \cdot n_a$ times for each training step, where S is the number of source domains and n_a is the number of search steps. These GGA training iterations only take place in the early stages of training and for a small percentage of the total training iterations (2% in our experiments). The rest of the iterations are vanilla ERM. Furthermore, inference is not affected by the application of GGA during training.

C. Full Experimental Results

In this section, we show detailed results of Table 2 in the main text. The results marked by †, ‡ are copied from Gulrajani and Lopez-Paz [19] and Wang *et al*. [48], respectively. Standard errors for the baseline methods are reported from three trials, if available from past literature. In green and red, we highlight the performance boost and decrease of applying **GGA** on top of each algorithm respectively, averaged over three trials. In addition, we also present detailed results for the DomainNet benchmark, without however including results for the combination of GGA with the baseline algorithms, due to computational restrictions. We also include experiments for the ColoredMNIST and RotatedMNIST datasets, where we reproduced the results for all baselines and report the average results over 5 runs. The below tables are better read in color.

When applying GGA to existing methods, the only difference regarding the baseline algorithm training is that "Algorithm 1" (i.e. GGA) is applied instead of the method's update rules for the duration of the annealing process (training steps A_s to A_e). The total epochs and method hyperparameters remain the same throughout training.

Table 1. Out-of-domain accuracies (%) on PACS.

Algorithm	A	C	P	S	Avg
IRM† [1]	84.8±1.3 (−3.6)	76.4±1.1 (+0.7)	96.7±0.6 (−0.3)	76.1±1.0 (−2.8)	83.5 (−1.5)
ERM‡ [43]	85.7 ±0.6	77.1 ±0.8	97.4 ±0.4	76.6 ±0.7	84.2
GroupDRO‡ [38]	83.5±0.9 (+3.6)	79.1±0.6 (+2.0)	96.7±0.7 (+0.5)	78.3±2.0 (−0.5)	84.4 (+1.4)
MTL‡[7]	87.5±0.8 (+1.3)	77.1±0.5 (+2.5)	96.4±0.8 (−0.8)	77.3±1.8 (+0.7)	84.6 (+0.9)
Mixup† [50]	86.1±0.5 (+1.2)	78.9±0.8 (−0.1)	97.6±0.1 (−0.4)	75.8±1.8 (+4.1)	84.6 (+1.2)
MMD‡ [30]	86.1±1.4 (+0.8)	79.4±0.9 (+0.5)	96.6±0.2 (−0.6)	76.5±0.5 (+3.7)	84.7 (+0.8)
VREx‡ [25]	86.0±1.6 (+0.2)	79.1±0.6 (−0.9)	96.9±0.5 (−0.1)	77.7±1.7 (+2.3)	84.9 (+0.5)
MLDG† [29]	85.5±1.4 (+1.1)	80.1±1.7 (+0.9)	97.4±0.3 (−0.9)	76.6±1.1 (+1.6)	84.9 (+0.7)
ARM‡ [55]	86.8±0.6 (−1.9)	76.8±0.5 (+4.7)	97.4±0.3 (−1.1)	79.3±1.2 (+0.4)	85.1 (+0.5)
Mixstyle‡ [58]	86.8±0.5 (+1.1)	79.0±1.4 (−0.3)	96.6±0.1 (−1.1)	78.5±2.3 (+1.4)	85.2 (+0.3)
CORAL† [41]	88.3±0.2 (−0.6)	80.0±0.5 (+1.3)	97.5±0.3 (+0.6)	78.8±1.3 (+1.5)	86.2 (+0.7)
SagNet† [34]	87.4±0.2 (−2.3)	80.7±0.5 (+1.0)	97.1±0.1 (−0.9)	80.0±1.0 (−1.8)	86.3 (−1.0)
RSC† [21]	85.4±0.9 (−1.8)	79.7±0.5 (+2.9)	97.6±0.9 (−1.0)	78.2±1.0 (+0.3)	85.2 (+0.1)
SAM ‡[18]	85.6±2.1 (+1.1)	80.9±1.2 (−0.8)	97.0±0.4 (−0.2)	79.6±1.6 (+1.2)	85.8 (+0.6)
GSAM ‡ [59]	86.9±0.1 (−0.4)	80.4±0.2 (+0.7)	97.5±0.0 (−1.1)	78.7±0.8 (+2.5)	85.9 (+0.4)
SAGM ‡ [48]	87.4±0.2 (+1.2)	80.2±0.3 (+1.1)	98.0±0.2 (−1.0)	80.8±0.6 (−0.4)	86.6 (+0.2)
GGA (ours)	88.8±0.2	80.1±0.3	97.3±0.2	81.2±0.5	87.3

Table 2. Out-of-domain accuracies (%) on VLCS.

Algorithm	C	L	S	V	Avg
GroupDRO‡ [38]	97.3±0.3 (+1.4)	63.4±0.9 (+1.7)	69.5±0.8 (+2.0)	76.7±0.7 (−2.9)	76.7 (+0.6)
MLDG† [29]	97.4±0.2 (+1.6)	65.2±0.7 (−0.4)	71.0±1.4 (+3.0)	75.3±1.0 (+0.9)	77.2 (+1.3)
MTL‡[7]	97.8±0.4 (+0.3)	64.3±0.3 (+1.8)	71.5±0.7 (+4.1)	75.3±1.7 (+1.6)	77.2 (+2.0)
ERM‡ [43]	98.0 ±0.3	64.7 ±1.2	71.4 ±1.2	75.2 ±1.6	77.3
Mixup† [50]	98.3±0.6 (+0.8)	64.8±1.0 (+1.7)	72.1±0.5 (+1.0)	74.3±0.8 (+3.6)	77.4 (+1.8)
MMD‡ [30]	97.7±0.1 (−0.9)	64.0±1.1 (+1.4)	72.8±0.2 (+1.4)	75.3±3.3 (+3.5)	77.5 (+1.3)
ARM‡ [55]	98.7±0.2 (−0.2)	63.6±0.7 (+2.2)	71.3±1.2 (+0.3)	76.7±0.6 (+1.4)	77.6 (+0.9)
SagNet† [34]	97.9±0.4 (−0.2)	64.5±0.5 (+1.7)	71.4±1.3 (+0.8)	77.5±0.5 (+1.4)	77.8 (+0.9)
Mixstyle‡ [58]	98.6±0.3 (−0.1)	64.5±1.1 (+1.9)	72.6±0.5 (+0.4)	75.7±1.7 (+0.4)	77.9 (+0.6)
VREx‡ [25]	98.4±0.3 (−0.9)	64.4±1.4 (+1.9)	74.1±0.4 (−1.7)	76.2±1.3 (+1.0)	78.3 (+0.1)
IRM† [1]	98.6±0.1 (−0.3)	64.9±0.9 (−3.5)	73.4±0.6 (+1.7)	77.3±0.9 (−1.5)	78.6 (−0.9)
CORAL† [41]	98.3±0.3 (+0.9)	66.1±0.6 (+1.8)	73.4±0.3 (−1.8)	77.5±1.0 (−2.1)	78.8 (−0.4)
RSC† [21]	97.9±0.1 (+0.6)	62.5±0.7 (+0.3)	72.3±1.2 (+0.4)	75.6±0.8 (−0.8)	77.1 (+0.2)
GSAM ‡ [59]	98.7±0.3 (+0.5)	64.9±0.2 (+0.5)	74.3±0.0 (+1.2)	78.5±0.8 (+1.8)	79.1 (+1.0)
SAM ‡[18]	99.1±0.2 (−0.2)	65.0±1.0 (+1.8)	73.7±1.0 (−0.2)	79.8±0.1 (+1.5)	79.4 (+0.7)
SAGM ‡ [48]	99.0±0.2 (−0.4)	65.2±0.4 (+0.5)	75.1±0.3 (−1.1)	80.7±0.8 (−0.2)	80.0 (−0.3)
GGA (ours)	99.1±0.2	67.5±0.6	75.1±0.3	78.0±0.1	79.9

Table 3. Out-of-domain accuracies (%) on OfficeHome.

Algorithm	A	C	P	R	Avg
Mixstyle‡ [58]	51.1±0.3 (+0.3)	53.2±0.4 (+0.7)	68.2±0.7 (+0.4)	69.2±0.6 (+0.6)	60.4 (+0.5)
IRM† [1]	58.9±2.3 (−3.5)	52.2±1.6 (−2.1)	72.1±2.9 (−1.7)	74.0±2.5 (−1.0)	64.3 (−2.1)
ARM‡ [55]	58.9±0.8 (+2.7)	51.0±0.5 (−0.3)	74.1±0.1 (+2.1)	75.2±0.3 (+3.4)	64.8 (+2.1)
GroupDRO‡ [38]	60.4±0.7 (+3.8)	52.7±1.0 (+1.2)	75.0±0.7 (+1.3)	76.0±0.7 (+2.2)	66.0 (+2.2)
MMD‡ [30]	60.4±0.2 (+3.1)	53.3±0.3 (−0.2)	74.3±0.1 (+3.1)	77.4±0.6 (+0.7)	66.4 (+1.6)
MTL‡[7]	61.5±0.7 (+0.8)	52.4±0.6 (−0.3)	74.9±0.4 (+1.0)	76.8±0.4 (+1.2)	66.4 (+0.4)
VREx‡ [25]	60.7±0.9 (+1.9)	53.0±0.9 (+0.8)	75.3±0.1 (+0.6)	76.6±0.5 (+0.2)	66.4 (+0.9)
ERM‡ [43]	63.1 ±0.3	51.9 ±0.4	77.2 ±0.5	78.1 ±0.2	67.6
MLDG† [29]	61.5±0.9 (+2.4)	53.2±0.6 (+0.2)	75.0±1.2 (+1.8)	77.5±0.4 (+0.6)	66.8 (+1.2)
Mixup† [50]	62.4±0.8 (+1.5)	54.8±0.6 (−1.7)	76.9±0.3 (+1.9)	78.3±0.2 (+0.4)	68.1 (+1.2)
SagNet† [34]	63.4±0.2 (+1.0)	54.8±0.4 (−1.9)	75.8±0.4 (+1.4)	78.3±0.3 (+0.7)	68.1 (+0.3)
CORAL† [41]	65.3±0.3 (+0.3)	54.4±0.6 (+0.2)	76.5±0.3 (−0.7)	78.4±1.0 (+1.0)	68.7 (+0.2)
RSC† [21]	60.7±1.4 (−1.1)	51.4±0.3 (+0.0)	74.8±1.1 (+0.6)	75.1±1.3 (+0.5)	65.5 (+0.0)
GSAM ‡ [59]	64.9±0.1 (−0.6)	55.2±0.2 (+1.1)	77.8±0.0 (+0.4)	79.2±0.2 (+0.3)	69.3 (+0.3)
SAM ‡[18]	64.5±0.3 (+0.7)	56.5±0.2 (+0.3)	77.4±0.1 (+1.0)	79.8±0.4 (+0.4)	69.6 (+0.6)
SAGM ‡ [48]	65.4±0.4 (−0.9)	57.0±0.3 (−0.8)	78.0±0.3 (+0.4)	80.0±0.2 (−1.1)	70.1 (−0.6)
GGA (ours)	64.3±0.1	54.4±0.2	76.5±0.3	78.9±0.2	68.5

Table 4. Out-of-domain accuracies (%) on TerraIncognita.

Algorithm	L100	L38	L43	L46	Avg
MMD‡ [30]	41.9±3.0 (+9.7)	34.8±1.0 (+9.8)	57.0±1.9 (+0.5)	35.2±1.8 (+5.9)	42.2 (+6.3)
GroupDRO‡ [38]	41.2±0.7 (−1.8)	38.6±2.1 (+9.4)	56.7±0.9 (±0.0)	36.4±2.1 (−1.6)	43.2 (+1.7)
Mixstyle‡ [58]	54.3±1.1 (−2.9)	34.1±1.1 (+8.8)	55.9±1.1 (−2.8)	31.7±2.1 (+2.9)	44.0 (+1.1)
ARM‡ [55]	49.3±0.7 (−3.0)	38.3±0.7 (+4.2)	55.8±0.8 (+2.0)	38.7±1.3 (−0.2)	45.5 (+0.8)
MTL‡[7]	49.3±1.2 (−5.9)	39.6±6.3 (+3.6)	55.6±1.1 (+2.1)	37.8±0.8 (+2.6)	45.6 (+0.9)
ERM‡ [43]	49.8 ±4.4	42.1 ±1.4	56.9 ±1.8	35.7 ±3.9	46.1
VREx‡ [25]	48.2±4.3 (+3.1)	41.7±1.3 (+0.7)	56.8±0.8 (+2.0)	38.7±3.1 (−0.4)	46.4 (+1.3)
IRM† [1]	54.6±1.3 (−4.3)	39.8±1.9 (−3.4)	56.2±1.8 (−3.8)	39.6±0.8 (−4.1)	47.6 (−3.9)
CORAL† [41]	51.6±2.4 (+3.1)	42.2±1.0 (−1.2)	57.0±1.0 (+1.1)	39.8±2.9 (−1.8)	47.6 (+0.3)
MLDG† [29]	54.2±3.0 (−2.5)	44.3±1.1 (+1.4)	55.6±0.3 (+5.1)	36.9±2.2 (+0.6)	47.8 (+1.2)
Mixup† [50]	59.6±2.0 (1.2)	42.2±1.4 (+7.6)	55.9±0.8 (+1.2)	33.9±1.4 (−0.9)	47.9 (+2.1)
SagNet† [34]	53.0±2.0 (+2.3)	43.0±1.4 (+0.2)	57.9±0.8 (−2.6)	40.4±1.4 (+2.9)	48.6 (+0.4)
SAM ‡[18]	46.3±1.0 (+3.3)	38.4±2.4 (+5.2)	54.0±1.0 (+1.9)	34.5±0.8 (−0.1)	43.3 (+2.6)
RSC† [21]	50.2±2.2 (−0.8)	39.2±1.4 (+1.0)	56.3±1.4 (+0.8)	40.8±0.6 (+0.2)	46.6 (+0.2)
GSAM ‡ [59]	50.8±0.1 (+3.8)	39.3±0.2 (+0.6)	59.6±0.0 (−2.2)	38.2±0.8 (+0.4)	47.0 (+0.6)
SAGM ‡ [48]	54.8±1.3 (±0.0)	41.4±0.8 (+6.3)	57.7±0.6 (−1.1)	41.3±0.4 (−5.5)	48.8 (−0.1)
GGA (ours)	55.9±0.1	45.5±0.1	59.7±0.1	41.5±0.1	50.6

Table 5. Out-of-domain accuracies (%) on DomainNet.

Algorithm	clip	info	paint	quick	real	sketch	Avg
MMD† [30]	32.1 ±13.3	11.0 ±4.6	26.8 ±11.3	8.7 ±2.1	32.7 ±13.8	28.9 ±11.9	23.4
GroupDRO† [38]	47.2 ±0.5	17.5 ±0.4	33.8 ±0.5	9.3 ±0.3	51.6 ±0.4	40.1 ±0.6	33.3
VREx† [25]	47.3 ±3.5	16.0 ±1.5	35.8 ±4.6	10.9 ±0.3	49.6 ±4.9	42.0 ±3.0	33.6
IRM† [1]	48.5 ±2.8	15.0 ±1.5	38.3 ±4.3	10.9 ±0.5	48.2 ±5.2	42.3 ±3.1	33.9
Mixstyle‡ [58]	51.9 ±0.4	13.3 ±0.2	37.0 ±0.5	12.3 ±0.1	46.1 ±0.3	43.4 ±0.4	34.0
ARM† [55]	49.7 ±0.3	16.3 ±0.5	40.9 ±1.1	9.4 ±0.1	53.4 ±0.4	43.5 ±0.4	35.5
Mixup‡ [50]	55.7±0.3	18.5±0.5	44.3±0.5	12.5±0.4	55.8±0.3	48.2±0.5	39.2
SagNet† [34]	57.7 ±0.3	19.0 ±0.2	45.3 ±0.3	12.7 ±0.5	58.1 ±0.5	48.8 ±0.2	40.3
MTL† [7]	57.9 ±0.5	18.5 ±0.4	46.0 ±0.1	12.5 ±0.1	59.5 ±0.3	49.2 ±0.1	40.6
MLDG† [29]	59.1 ±0.2	19.1 ±0.3	45.8 ±0.7	13.4 ±0.3	59.6 ±0.2	50.2 ±0.4	41.2
CORAL† [41]	59.2 ±0.1	19.7 ±0.2	46.6 ±0.3	13.4 ±0.4	59.8 ±0.2	50.1 ±0.6	41.5
ERM‡ [43]	63.0 ±0.2	21.2 ±0.2	50.1 ±0.4	13.9 ±0.5	63.7 ±0.2	52.0 ±0.5	43.8
RSC† [21]	55.0±1.2	18.3±0.5	44.4±0.6	12.2±0.2	55.7±0.7	47.8±0.9	38.9
SAM‡ [18]	64.5±0.3	20.7±0.2	50.2±0.1	15.1±0.3	62.6±0.2	52.7±0.3	44.3
GSAM ‡ [59]	64.2±0.3	20.8±0.2	50.9±0.0	14.4±0.8	63.5±0.2	53.9±0.2	44.6
SAGM ‡ [48]	64.9±0.2	21.1±0.3	51.5±0.2	14.8±0.2	64.1±0.2	53.6±0.2	45.0
GGA (ours)	64.0±0.2	22.2±0.3	51.7±0.1	14.3±0.2	64.1±0.4	54.3±0.3	45.2

Table 6. Out-of-domain accuracies (%) on ColoredMNIST (left) and RotatedMNIST (right).

Algorithm	0.1	0.2	0.9	Avg	0	15	30	45	60	75	Avg
IRM‡ [1]	56.8±4.5	63.5±2.7	10.2±0.2	43.5	95.5±0.4	98.7±0.2	98.7±0.1	98.5±0.3	98.7±0.1	96.1±0.1	97.7
MLDG [29]	71.5±0.6	73.0±0.1	10.1±0.2	51.5	94.7±0.7	98.8±0.1	98.8±0.1	98.8±0.1	98.7±0.1	95.9±0.4	97.6
MTL [7]	71.3±0.6	72.9±0.2	10.2±0.1	51.5	94.6±1.1	98.6±0.2	98.8±0.1	98.7±0.1	98.7±0.3	95.3±0.7	97.4
Mixup [50]	71.5±0.8	73.2±0.3	10.2±0.2	51.6	94.9±0.5	98.8±0.1	98.8±0.2	98.8±0.1	98.8±0.1	95.7±0.5	97.6
SagNet [34]	72.0±0.5	72.8±0.5	9.9±0.3	51.6	95.5±0.3	98.9±0.1	99.0±0.1	98.8±0.2	98.8±0.1	95.9±0.4	97.8
ERM [43]	71.8±0.9	73.3±0.4	9.9±0.3	51.7	95.1±0.6	98.7±0.2	98.7±0.2	98.7±0.2	98.8±0.1	95.6±0.4	97.6
ARM [55]	74.5±3.8	71.1±1.8	9.9±0.3	51.8	95.1±1.1	98.8±0.2	98.8±0.1	98.8±0.1	98.8±0.1	96±0.6	97.7
CORAL [41]	72.3±0.7	72.8±0.4	10.5±0.3	51.8	95.6±0.3	98.9±0.1	98.9±0.1	99.0±0.0	98.9±0.1	96.1±0.3	97.9
Fish [40]	71.7±0.5	73.2±0.5	10.4±0.2	51.8	95.3±0.6	98.9±0.1	98.9±0.2	98.9±0.1	98.9±0.1	95.6±0.6	97.7
GroupDRO [38]	72.6±0.6	73.5±0.4	9.9±0.2	52.0	95.9±0.6	98.7±0.2	98.6±0.1	98.7±0.1	98.7±0.1	96.0±0.2	97.8
VREx [25]	72.9±0.3	72.9±0.4	10.3±0.6	52.0	95.7±0.6	98.9±0.2	98.7±0.1	98.9±0.2	98.9±0.1	95.8±0.4	97.8
SAM [18]	71.1±0.5	73.3±0.4	10.1±0.3	51.5	95.7±0.2	99.0±0.1	98.9±0.1	98.9±0.1	98.9±0.1	96.2±0.4	97.9
GSAM [59]	71.8±0.3	73.2±0.2	9.9±0.2	51.6	94.9±0.1	98.9±0.1	98.9±0.2	99.0±0.2	98.8±0.1	96.0±0.1	97.7
RSC [21]	72.5±0.3	72.4±0.6	10.2±0.5	51.7	94.2±1.1	98.6±0.1	98.7±0.2	98.6±0.2	98.7±0.2	95.7±0.7	97.4
SAGM [48]	71.5±0.8	73.6±0.5	10.6±0.6	51.9	95.4±0.4	98.9±0.1	98.9±0.1	98.9±0.1	98.9±0.1	95.9±0.5	97.8
GGA (ours)	71.2±0.7	73.1±0.6	11.5±0.4	51.9	95.1±0.8	99.0±0.1	99.0±0.3	98.8±0.1	98.8±0.2	96.1±0.4	97.8

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 4, 5, 6

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 2

[3] Aristotelis Ballas and Christos Diou. Towards domain generalization for ecg and eeg classification: Algorithms and benchmarks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(1):44–54, 2024. 1

[4] Aristotelis Ballas and Christos Diou. Multi-scale and multi-layer contrastive learning for domain generalization. *IEEE Transactions on Artificial Intelligence*, 2024. 2

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 2

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1

[7] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021. 4, 5, 6

[8] Francesco Cappio Borlino, Antonio D'Innocente, and Tatiana Tommasi. Rethinking domain generalization baselines. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9227–9233. IEEE, 2021. 1

[9] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 2

[10] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. 2

[11] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 301–318. Springer, 2020. 1

[12] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 2

[13] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10312–10322, 2023. 1

[14] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313, 2017. 2

[15] Theekshana Dissanayake, Tharindu Fernando, Simon Denman, Houman Ghaemmaghami, Sridha Sridharan, and Clinton Fookes. Domain generalization in biosignal classification. *IEEE Transactions on Biomedical Engineering*, 68(6):1978–1989, 2021. 1

[16] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. Adarnn: Adaptive learning and forecasting of time series. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 402–411, 2021. 1

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[18] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 2, 4, 5, 6

[19] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 3

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 1

[21] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *European Conference on Computer Vision*, 2020. 2, 4, 5, 6

[22] Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023. 1

[23] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational

autoencoders. In *Medical Imaging with Deep Learning*, pages 322–348. PMLR, 2020. 2

[24] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *International Conference on Computer Vision*, 2021. 2

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. 4, 5, 6

[26] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. 2

[27] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pages 5639–5650. PMLR, 2020. 2

[28] Kimin Lee, Kibok Lee, Jinwoo Shin, and Honglak Lee. Network randomization: A simple technique for generalization in deep reinforcement learning. *arXiv preprint arXiv:1910.05396*, 2019. 2

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1, 2, 4, 5, 6

[30] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Computer Vision and Pattern Recognition*, 2018. 4, 5, 6

[31] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018. 2

[32] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pages 7313–7324. PMLR, 2021. 2

[33] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11749–11756, 2020. 2

[34] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Computer Vision and Pattern Recognition*, 2021. 4, 5, 6

[35] Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in bioengineering and biotechnology*, 7:198, 2019. 1

[36] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International conference on machine learning*, pages 7728–7738. PMLR, 2020. 1

[37] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020. 1

[38] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 4, 5, 6

[39] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018. 1

[40] Yuge Shi, Jeffrey Seely, Philip Torr, Siddharth N, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022. 2, 6

[41] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016. 4, 5, 6

[42] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 2

[43] V Vapnik. Statistical learning theory. *NY: Wiley*, 1998. 4, 5, 6

[44] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2019. 1

[45] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 1

[46] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35 (8):8052–8072, 2022. 1

[47] Jun Wang, He Ren, Changqing Shen, Weiguo Huang, and Zhongkui Zhu. Multi-scale style generative and adversarial contrastive networks for single domain generalization fault diagnosis. *Reliability Engineering & System Safety*, 243:109879, 2024. 1

[48] 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*, pages 3769–3778, 2023. 2, 3, 4, 5, 6

[49] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12):4237–4248, 2020. 2

[50] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI Conference on Artificial Intelligence*, 2020. 1, 4, 5, 6

[51] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 2

[52] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2100–2110, 2019. 1

[53] 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, 2022. 1

[54] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020. 1

[55] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv preprint arXiv:2007.02931*, 2020. 2, 4, 5, 6

[56] Huailiang Zheng, Yuantao Yang, Jiancheng Yin, Yuqing Li, Rixin Wang, and Minqiang Xu. Deep domain generalization combining a priori diagnosis knowledge toward cross-domain fault diagnosis of rolling bearing. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021. 1

[57] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. 2

[58] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 1, 4, 5, 6

[59] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C Dvornek, sekhar tatikonda, James s Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022. 2, 4, 5, 6