

Domain Generalization Guided by Gradient Signal to Noise Ratio of Parameters

Mateusz Michalkiewicz
University of Queensland

Masoud Faraki
NEC Labs America

Xiang Yu
Amazon *

Manmohan Chandraker
NEC Labs America,
University of California, San Diego

Mahsa Baktashmotlagh
University of Queensland

Abstract

Overfitting to the source domain is a common issue in gradient-based training of deep neural networks. To compensate for the over-parameterized models, numerous regularization techniques have been introduced such as those based on dropout. While these methods achieve significant improvements on classical benchmarks such as ImageNet, their performance diminishes with the introduction of domain shift in the test set i.e. when the unseen data comes from a significantly different distribution. In this paper, we move away from the classical approach of Bernoulli sampled dropout mask construction and propose to base the selection on gradient-signal-to-noise ratio (GSNR) of network’s parameters. Specifically, at each training step, parameters with high GSNR will be discarded. Furthermore, we alleviate the burden of manually searching for the optimal dropout ratio by leveraging a meta-learning approach. We evaluate our method on standard domain generalization benchmarks and achieve competitive results on classification and face anti-spoofing problems.

1. Introduction

In recent years, deep neural networks achieved remarkably good results on several classification tasks, facilitated by regularization methods that successfully reduce overfitting of large models to the training data. A simple yet powerful technique is *Dropout* [51], which mutes randomly chosen activations of fully connected layers at each training iteration. A better-suited variant for Convolutional Neural Networks is *Dropblock* [14], which masks contiguous regions of feature maps with spatial correlation. While existing regularization techniques achieve great results,

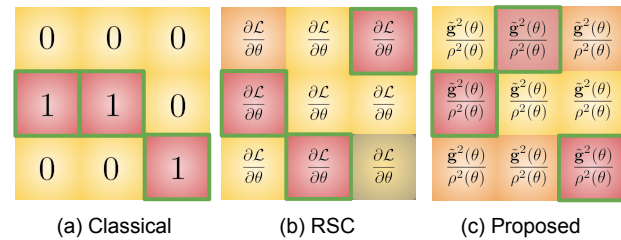


Figure 1: Different approaches to dropout mask construction. Left to right: (a) classical approach where values are sampled from a Bernoulli distribution [51]; (b) RSC [19], where parameters with highest gradients are masked ; (c) ours, where parameters with highest gradient-signal-to-noise ratio are masked. The amount of discarded parameters is controlled by a dropout ratio p which is typically manually chosen, whereas in our approach it is automatically learned.

their success is based on the underlying assumption that the train and test data follow similar distributions. A more practical scenario, however, is presented in a *domain generalization* setting, where there is a distribution shift between the train and test set [50, 74]. Here, models equipped with classical regularization techniques often fail to generalize their inference to unseen examples.

The goal of our work is to build a model that is robust to the domain shift and performs equally well on both source and unseen test domains. We build our model on two observations. First, models with high Gradient Signal to Noise Ratio (GSNR), defined as the ratio of squared mean over variance of parameters gradients on a particular data distribution, exhibit a smaller generalization gap *i.e.* their performance does not drastically decrease when evaluated on the unseen data [32]. Second, by iteratively dropping the most predictive parameters, the model is forced to learn less dom-

*Work done while Xiang was at NEC Labs America

inant features which might correspond to domain-invariant features, thus improving performance on unseen domains [19].

In light of this discussion, we carefully design a dropout strategy to drop parameters with highest gradient-signal-to-noise ratio in each training step which we illustrate in Figure 1. As a result, the overall GSNR of the model improves which leads to a better generalization performance. Furthermore, we have observed that different blocks of neural network favour different dropout ratios. Therefore, we replace the standard approach of applying a fixed and manually chosen dropout probability p by leveraging a learning-to-learn technique [10, 30] to learn the dropout probability for each neural network block. Lastly, we validate our approach through extensive experiments on benchmark domain generalization datasets on classification and face recognition tasks, and show that our approach outperforms all the baselines including the ones based on a standard dropout strategy. More specifically, our GSNR-guided dropout is complementary to the recent method of Representation Self-Challenging (RSC) [19] which drops out features based on gradient magnitudes with a fixed probability. We perform extensive experiments on widely used domain generalization datasets *i.e.*, DomainNet [40], OfficeHome [54] and PACS [26] and the OCIM benchmark for face anti-spoofing consisting of 4 different datasets: OULU-NPU [2], CASIA-FASD [70], MSU-MFSD [6], and REPLAY-ATTACK [61]. The results show that our method consistently leads to improvements compared to the existing methods.

To summarize, our contributions are three-fold:

1. We introduce a novel dropout strategy based on GSNR, that can be easily incorporated in any standard convolutional neural network architecture.
2. We alleviate the problem of choosing optimal dropout ratios through a novel meta-learning framework.
3. We empirically validate our approach on a number of domain generalization benchmark datasets for object classification and face anti-spoofing tasks.

2. Related Work

2.1. Domain Generalization

Domain Generalization (DG) algorithms seek to generalize beyond several source domains while reducing the distributional shift to target domains. The goal is to learn a model that is generalizable to any unseen domain. DG has been studied from a number of different angles.

A straightforward approach proposed by [53], is to utilize the empirical risk minimization strategy and train the DG model with multiple source domains. This standard idea which is borrowed from the fully supervised learning

strategy is also effective for domain generalization. Another very common approach is domain alignment where the goal is to extract a domain invariant feature representation [37, 9].

Another approach is by decoupling domain agnostic features from domain specific ones as done in [26]. To boost generalization ability, Zhou *et al.* [74] suggest that a style transfer strategy captured by the bottom layers of a network, be applied to the feature statistics. This is motivated by the fact that visual domains are mostly related to image styles.

Patch shuffling input images [39] or parsing them through random convolutions [64] are one of many examples of data augmentation approaches to DG. Domain shift may also be handled by training a separate classifier for each source domain and using the one that best fits the target domain [63].

Learning-to-learn techniques [10, 30, 31] emulate the domain shift during training by constructing meta-train and meta-test sets, assuming that good performance on meta-test set will translate well to unseen domains. An extension is suggested in [68] to adapt the empirical risk minimization loss to a context model with a learning to learn method.

2.2. Dropout Regularization

A commonly used regularization technique to prevent overfitting is Dropout [51], where a binary mask is sampled from a Bernoulli distribution at each training iteration and applied to mute random parameters. Further, to compensate for the masked parameters, the outputs are scaled by $\frac{1}{1-p}$, where p is the probability that a given parameter is muted. As a result, networks can learn alternative pathways to improve their predictive abilities on unseen data.

This simple procedure inspired numerous other dropout-based approaches. Popular examples are: DropPath [24] where an entire layer is muted, SpatialDropout [52] where dropout is applied channel-wise, CutOut [8] where random patches of input images are dropped, or AlphaDropout [22] which is also applied to input images while preserving the original mean and standard deviation. An approach closely related to our method is DropBlock [14] which drops out square patches of feature maps of a given size and is better-suited than Dropout for Convolutional Neural Networks as it is more efficient in muting semantic information.

These techniques showed improvement in the generalization capabilities of neural networks as long as the train and test examples follow the same distribution. However, they are no longer effective when evaluated under a domain generalization setup where a significant domain shift is introduced [75]. Here, zeroing the most predictive parts of features maps, such as those with highest activation values or gradients [19], has yielded better results.

3. Methodology

3.1. Notations

In the next sections we will use the following notations. Let x and y denote images and their corresponding labels sampled from a data distribution \mathcal{Z} . Let $f_\theta(\mathbf{x}, M)$ denote a modified ResNet18 network [16] parameterized by θ , where a tensor M is applied to each activation block. Let $m_j^{(i)}$ denote j -th element of matrix $M^{(i)}$. Finally, let $\text{sort}(\cdot)_k$ and $\text{top } k(\cdot, k)$ denote k largest elements and k -th largest element respectively.

3.2. Background

Understanding why neural networks generalize is a fundamental problem. A number of works study the generalization ability of gradient-based methods [41, 66, 28]. Recent works introduce the concept of network stiffness, where alignment of loss gradients is linked to generalization [11]. More recently, two metrics are proposed to study the generalization gap [28]: gradient-signal-to-noise ratio (GSNR) and one-step-generalization ratio (OSGR). Given a network f , loss function \mathcal{L} , images x , and their corresponding labels y sampled from a data distribution \mathcal{Z} , the GSNR of a parameter θ is defined as the ratio between parameters' mean gradients and the corresponding variance, with respect to the loss function, as below:

$$r(\theta) = \frac{\tilde{\mathbf{g}}^2(\theta)}{\rho^2(\theta)}, \quad (1)$$

where:

$$\begin{aligned} \tilde{\mathbf{g}}(\theta) &= E_{(x,y) \sim \mathcal{Z}} \left(\frac{\partial \mathcal{L}(f(x, \theta), y)}{\partial \theta} \right), \\ \rho^2(\theta) &= \text{Var}_{(x,y) \sim \mathcal{Z}} \left(\frac{\partial \mathcal{L}(f(x, \theta), y)}{\partial \theta} \right). \end{aligned} \quad (2)$$

Given an empirical training loss $L[D]$ on dataset D and an empirical testing loss $L[D']$ on dataset D' :

$$\begin{aligned} L[D] &= \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i, \theta), y), \\ L[D'] &= \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x'_i, \theta), y'), \end{aligned} \quad (3)$$

we can define OSGR as:

$$\mathbf{R}(\mathcal{Z}, n) = \frac{E_{D, D' \sim \mathcal{Z}^n}(\Delta L[D'])}{E_{D \sim \mathcal{Z}^n}(\Delta L[D])}, \quad (4)$$

where n denotes the size of datasets D and D' , respectively. Note that OSGR of 1 means that the performance gap between train set and test set is 0, which indicates perfect generalization. Following [32] we can rewrite $\mathbf{R}(\mathcal{Z}, n)$ as:

$$\mathbf{R}(\mathcal{Z}, n) = 1 - \frac{\sum_j \rho_j^2}{n \sum_j (\frac{1}{n} \rho_j^2 + \tilde{\mathbf{g}}_j^2)}, \quad (5)$$

which shows OSGR's dependence on the magnitude of the gradients of network's parameters $\tilde{\mathbf{g}}$ and their corresponding standard deviations ρ . Assuming the learning rate is small, OSGR can be expressed as:

$$\mathbf{R}(\mathcal{Z}, n) = 1 - \frac{1}{n} \sum_j W_j \frac{1}{r_j + \frac{1}{n}}, \quad (6)$$

where the weights $W_j = \frac{E_{D \sim \mathcal{Z}^n}(\Delta L_j[D])}{E_{D \sim \mathcal{Z}^n}(\Delta L[D])}$ sum up to 1. This shows that high GSNR of a model correlates with high OSGR, *i.e.*, strong generalization. However, no clear way of enhancing the model's GSNR is presented.

3.3. Proposed Approach

The goal of domain generalization is to design a model that would demonstrate good generalization capabilities towards unseen domains. One way of measuring the generalization ability of a neural network is by looking at their gradient-signal-to-noise ratio (GSNR). In [32], a quantitative relationship is developed between parameter's gradient-signal-to-noise ratio and the ability of the model to generalize well to unseen examples. Since the proposed ratio between mean and variance of the gradients cannot be readily optimised, we propose to enhance the model's GSNR by iteratively zeroing parameters that exhibit high GSNRs.

As noted in [72], simply muting random parameters during training is an ineffective strategy in a domain generalization setup. However, muting the most significant features can indeed improve cross-domain generalization. The importance of the most predictive features can be measured by looking at the magnitude of the gradients, as studied in [19]. We instead focus on the parameters with high GSNR and combine it with the DropBlock technique to improve generalization capabilities and outperform [19], as well as domain generalization baselines. In Section 3.3.1, we describe the proposed procedure in detail.

We also observe that the optimal dropout ratio varies across different ResNet blocks and different domains. A trivial solution is to find those ratios through parameter grid search. This approach, however, can be infeasible as the number of possible configurations grows exponentially with increasing depth of the network. To alleviate this computational burden, alternative dropout strategies have been proposed [19] which are only applied to a single ResNet block. To improve on the aforementioned sub-optimal strategies, we propose to learn parameters which modulate the amount of activations to be muted through a meta-learner. We describe this approach in Section 3.3.2.

Finally, a comparative analysis in Section 3.3.3 highlights the differences between our approach and the one of [19].

3.3.1 GSNR-Guided DropBlock

Our approach follows a dropout regularization procedure where, in each training step, we construct a dropout mask that mutes a subset of activations. Concretely, the forward pass consists of the following three steps:

1. First, we calculate the gradients of the loss function with respect to the parameters of i -th ResNet block.

$$\mathbf{g}_i^{(1)} = \frac{\partial \mathcal{L}(f_\theta(\mathbf{x}, \mathbf{1}), \mathbf{y})}{\partial \theta_{\text{block}_i}}. \quad (7)$$

At this stage, an identity tensor $\mathbf{1}$ is used to compute the logits. Then, we calculate the gradient-signal-to-noise ratio for each parameter θ_j :

$$r_j = \frac{\mathbb{E}_{(x,y) \sim \mathcal{Z}}(\mathbf{g}(x, y, \theta_j))}{\text{Var}_{(x,y) \sim \mathcal{Z}}(\mathbf{g}(x, y, \theta_j))}. \quad (8)$$

Here, the mean and variance of the data distribution \mathcal{Z} is approximated by the mean and variance within the current batch. Then, we construct a binary mask where parameters with largest GSNR are zeroed:

$$m_j^{(1)} = \begin{cases} 1, & \text{if } r_j \geq \tau \\ 0, & \text{otherwise} \end{cases}, \quad (9)$$

with the threshold τ being the k -th largest GSNR value in ResNet block i :

$$\tau_i = \text{top}_k(r_j, k). \quad (10)$$

Following common practice, we design a mask identifying whether muting the parameter should occur:

$$m_j^{(2)} \sim \text{Bernoulli}(p_{\text{gsnr}}). \quad (11)$$

2. We combine $M_j^{(1)}$ and $M_j^{(2)}$ to select which activations should be set to 0:

$$M = M^{(1)} \times M^{(2)}. \quad (12)$$

3. Finally, we compute the gradients of the loss function with respect to all parameters:

$$\mathbf{g} = \frac{\partial \mathcal{L}(f_\theta(\mathbf{x}, M), \mathbf{y})}{\partial \theta}, \quad (13)$$

and feed them to the optimizer.

Our approach is summarized in Algorithm 1. As described earlier, this procedure requires manually selecting the dropout ratio, (equation 10) which can be prohibitively expensive as different ResNet blocks and domains favour different dropout ratios. To alleviate this problem, in the next section, we propose to use a meta-learning approach for learning the dropout ratios.

Algorithm 1 DropBlock with GSNR

- 1: **Input** : activations A , $block_size$, γ , $mode$, gradients $g = \nabla_A(\mathcal{L})$, k , p
 - 2: **if** $mode == \text{Inference}$ **then**
 - 3: return A
 - 4: **end if**
 - 5: Compute mask M
 - a. Compute GSNR with eq. 8
 - b. Compute threshold τ with eq. 10
 - c. Compute mask $M^{(1)}$ using eq. 9
 - d. Randomly sample $M^{(2)}$ using eq. 11
 - e. Combine both masks: $M = M^{(1)} \times M^{(2)}$
 - 6: For each zero position $M_{i,j}$, create a spatial square mask with the center being $M_{i,j}$, the width, height being $block_size$ and set all the values of M to be zero
 - 7: **Apply**: $A = A \times M$
 - 8: Normalize the features:

$$A = A \times \text{count}(M) / \text{count_ones}(M)$$
-

3.3.2 Meta-learning the Dropout Ratios

In this section, we aim to bypass the burden of the exhaustive grid search by employing learning-to-learn technique. However, our procedure described in Algorithm 1 cannot be readily incorporated into a meta-learning framework: the sorting operation is not differentiable and computing $\frac{\partial \mathcal{L}}{\partial k}$ which is needed for the backward pass, cannot be handled through auto-differentiation engines (e.g., in Pytorch).

To alleviate this problem, we propose to sample the $M^{(2)}$ mask from a uniform distribution and apply a scaled hard sigmoid function ϕ :

$$m_j^{(2)} \sim \mathcal{U}[-1, 1], \quad (14)$$

$$M^{(2)} = \phi(M^{(2)} + p), \quad (15)$$

where ϕ is defined as:

$$\phi(x) = \begin{cases} 0, & \text{if } x \leq -3 \\ 1, & \text{if } x \geq 3 \\ \frac{x}{6} + \frac{1}{2}, & \text{otherwise} \end{cases}. \quad (16)$$

Note, that the parameter p modulates the amount of activations to be zeroed, *i.e.* describes the dropout ratio, and does not pose any issues with differentiation. Our modified procedure can now be used in a meta-learning framework which we base on [30]. Concretely, during each training step, we randomly pick a subset of the current batch B_i as the meta-training set \mathcal{D}_{mtr} . We adapt a meta-testing set construction from [31], where we sample k examples with the largest distance (ℓ_2 norm between logits) from \mathcal{D}_{mtr} :

$$\mathcal{D}_{\text{mte}} = \{x_m \in B_i \setminus \mathcal{D}_{\text{mtr}} \mid \text{sort}(\max_{x_j \in \mathcal{D}_{\text{mtr}}} \|f_\theta(x_m, \mathbf{1}) - f_\theta(x_j, \mathbf{1})\|_2)_k\} \quad (17)$$

The meta-learning pass consists of two steps: a meta-train step and a meta-test step. To adapt the learner to the classification task, we need to compute the gradients with respect to all ResNet blocks first:

$$\mathcal{L}_{\mathcal{D}_{\text{mtr}}} = \frac{1}{|\mathcal{D}_{\text{mtr}}|} \sum_{(x,y) \in \mathcal{D}_{\text{mtr}}} \mathcal{L}(f_{\theta}(\mathbf{x}, \mathbb{1}), \mathbf{y}), \quad (18)$$

$$\mathbf{g}_i^{(1)} = \frac{\partial \mathcal{L}_{\text{mtr}}(f_{\theta}(\mathbf{x}, \mathbb{1}), \mathbf{y})}{\partial \theta_{\text{block}_i}}. \quad (19)$$

We then construct the dropout mask using previously described equations 9 and 15 and use it to compute the meta-training loss and update the learners weights:

$$\mathcal{L}_{\mathcal{D}_{\text{mtr}}} = \frac{1}{|\mathcal{D}_{\text{mtr}}|} \sum_{(x,y) \in \mathcal{D}_{\text{mtr}}} \mathcal{L}(f_{\theta}(\mathbf{x}, M), \mathbf{y}),$$

Algorithm 2 Meta GSNR

```

1: Input :  $\beta, \gamma, k$ , images, labels,  $f_{\theta}(x, M)$ 
2: Initialize  $\alpha, \rho$ 
3: for each epoch do
4:   for each batch  $B_i$  do
5:      $\mathcal{D}_{\text{mtr}} \leftarrow$  random subset of  $B_i$ 
6:     Meta-train:
7:      $\mathcal{L}_{\mathcal{D}_{\text{mtr}}} = \frac{1}{|\mathcal{D}_{\text{mtr}}|} \sum_{(x,y) \in \mathcal{D}_{\text{mtr}}} \mathcal{L}(f_{\theta}(\mathbf{x}, \mathbb{1}), \mathbf{y})$ 
8:     for each block  $l$  do
9:        $\mathbf{g}_l^{(1)} = \frac{\partial \mathcal{L}_{\text{mtr}}(f_{\theta}(\mathbf{x}, \mathbb{1}), \mathbf{y})}{\partial \theta_{\text{block}_l}}$ 
10:       $r_{l_j} = \frac{\text{mean}_{x,y \in \mathcal{D}_{\text{mtr}}} g_l(x,y, \theta_{l_j})}{\text{var}_{x,y \in \mathcal{D}_{\text{mtr}}} g_l(x,y, \theta_{l_j})}$ 
11:       $\tau_l = \text{top } k(r_{l_j}, k)_j$ 
12:       $m_{l_j}^{(1)} = \begin{cases} 1, & \text{if } r_{l_j} \geq \tau_l \\ 0, & \text{otherwise} \end{cases}$ 
13:       $m_{l_j}^{(2)} \sim \mathcal{U}[-1, 1]$ 
14:       $M_l^{(2)} = \phi(M^{(2)} + p)$ 
15:       $M_l = M_l^{(1)} \times M_l^{(2)}$ 
16:     end for
17:      $\mathcal{L}_{\mathcal{D}_{\text{mtr}}} = \frac{1}{|\mathcal{D}_{\text{mtr}}|} \sum_{(x,y) \in \mathcal{D}_{\text{mtr}}} \mathcal{L}(f_{\theta}(\mathbf{x}, M), \mathbf{y})$ 
18:      $\mathbf{g} = \nabla_{\theta} \mathcal{L}_{\mathcal{D}_{\text{mtr}}}$ 
19:      $\theta'_i = \theta - \alpha \circ \mathbf{g}$ 
20:     Meta-test:
21:     Construct  $\mathcal{D}_{\text{mte}}$  using eq. 17
22:      $\mathcal{L}_{\mathcal{D}_{\text{mte}}} = \frac{1}{|\mathcal{D}_{\text{mte}}|} \sum_{(x,y) \in \mathcal{D}_{\text{mte}}} \mathcal{L}(f_{\theta'}(\mathbf{x}, M), \mathbf{y})$ 
23:      $\mathcal{L}_i = \gamma \mathcal{L}_{\mathcal{D}_{\text{mtr}}} + (1 - \gamma) \mathcal{L}_{\mathcal{D}_{\text{mte}}}$ 
24:   end for
25:   Meta-optimization:
26:    $(\theta, \alpha, \rho) \leftarrow (\theta, \alpha, \rho) - \beta \nabla_{\theta, \alpha, \rho} \sum_i \mathcal{L}_i$ 
27: end for

```

$$\theta' = \theta - \alpha \circ \mathbf{g}^{(1)}.$$

We calculate the meta-test loss using the updated learners weight θ' :

$$\mathcal{L}_{\mathcal{D}_{\text{mte}}} = \frac{1}{|\mathcal{D}_{\text{mte}}|} \sum_{(x,y) \in \mathcal{D}_{\text{mte}}} \mathcal{L}(f_{\theta'}(\mathbf{x}, \mathbf{g}), \mathbf{y}). \quad (20)$$

Finally, the meta-train and meta-test losses are combined with a γ -weighted average:

$$\mathcal{L}_i = \gamma \mathcal{L}_{\mathcal{D}_{\text{mtr}}} + (1 - \gamma) \mathcal{L}_{\mathcal{D}_{\text{mte}}}, \quad (21)$$

with the parameters θ , learning rates α and dropout ratios p being updated through the adaptation steps:

$$(\theta, \alpha, \rho) \leftarrow (\theta, \alpha, \rho) - \beta \nabla_{\theta, \alpha, \rho} \sum_i \mathcal{L}_i. \quad (22)$$

A summary of the procedure is in Algorithm 2.

3.3.3 Comparison with RSC

Representation Self-Challenging (RSC) [19] is the closest work in the literature to our approach. RSC drops the features associated with higher gradients at each iteration. It introduces a number of extra parameters that needs tuning, such as: a parameter controlling whether the features are first pooled spatial-wise or channel-wise, or both, and the percentage of input images affected by RSC in each batch. Additionally, the data-specific dropout ratio p is manually selected and only applied to the last ResNet block.

The fundamental difference between our approach and the one of RSC is the metric that is used to determine the importance of the features. While RSC looks at the magnitude of the gradients, our approach measures the gradient-signal-to-noise of the network's parameters. Moreover, our approach is much more portable as it simply overwrites the dropout mask construction in the DropBlock procedure. Lastly, in our approach, the dropout ratio p is learned and applied to all ResNet blocks. Note that, learning dropout ratio in GSNR-guided dropblock allows our approach to achieve better performance than RSC.

4. Experiments

4.1. Experimental Setup

In all of our experiments, we follow a leave-one-out evaluation protocol where we train our model on three source domains and evaluate on the fourth one. We use ResNet18 network [16] as backbone which we pretrain on the ImageNet [7]. We base our learning-to-learn framework on Meta-SGD. Our dropout procedure is based on DropBlock [14], where we set the block size to 3. Our hyperparameters

γ and β are set to 0.5 and 0.001, respectively. Following [19], we use SGD optimizer with weight decay 0.0005 and momentum 0.9 and we initialize the learning rate with 0.004 and decay it by 0.1 after 24 epochs. Our batch size was set to 128. In classification experiments, we have used an identical codebase of [19] and used the default 30 epochs for all dropout-based methods (ours, RSC, SpatialDropout, etc). To be consistent with the FAS competitors, we have trained our SSAN+ours model for 1200 epochs (as in done [59]) and our EPCR+ours model for 800 epochs (as done in [60]).

Unless noted otherwise, we follow a common practice of stacking a standard set of data augmentations. That is, we randomly apply the following procedures: crop and resize, horizontal flip, color jitter (*i.e.* brightness, contrast, saturation, and hue), grayscale, and random convolution.

4.2. PACS Classification

PACS [26] is a popular domain generalization benchmark dataset where images are gathered in four different domains of: *art painting*, *cartoon*, *sketch*, and *photo*, with a challenging domain shift. Each image belongs to one of the seven seen categories: *dog*, *elephant*, *giraffe*, *guitar*, *horse*, *house*, and *person*. In total, PACS contains close to 10,000 images with over 1500 falling into each domain. Following the original setup, we split the images from training domains to 90% train - 10% val and test on the whole held-out domain.

Results gathered in Table 1 show that our proposed method outperforms our main competitor [19] on all 4 domains and improves the average performance by almost 2%. We further demonstrate the applicability of our approach using two other standard architectures: AlexNet [23] and ResNet-50 [16]. The results can be found in the supplementary materials.

4.3. Office-Home Classification

Office-Home [54] is another popular benchmark dataset containing four domains: *art*, *clipart*, *product*, and *real-world*. It consists of approximately 15,000 images grouped into 65 unique categories, representing objects commonly found in household and office environments. Each of those categories is built of approximately 70 images. We followed the standard practice of training on all images of training domains and testing on the held out one. Similar to PACS evaluation setup, our model improves the average performance of RSC by over 2% on all domains which we present in Table 2.

4.4. miniDomainNet Classification

Compared to PACS and Office-Home, DomainNet [40] is a significantly larger dataset spanning over 6 domains: *clipart*, *infograph*, *painting*, *quickdraw*, *real*, and *sketch*. A

Table 1: Classification accuracy (%) on the PACS dataset [26]. The bold numbers indicate the best performance averaged across all domains, second best is underlined.

PACS	artpaint	cartoon	sketch	photo	Avg \uparrow
Deep All [15]	78.63	75.27	68.72	96.08	79.68
DMG [4]	76.90	80.38	75.21	93.35	81.46
MMLD [35]	81.28	77.16	72.29	96.09	81.83
L2A-OT [71]	83.30	78.20	73.60	96.20	82.80
DSON [45]	84.67	77.65	82.23	95.87	85.11
MixStyle [72]	84.10	78.80	75.90	96.10	83.70
LDSGD [58]	81.44	79.56	80.58	95.51	84.27
NAS-OoD [1]	83.74	79.69	77.27	96.23	84.23
SFA-A [29]	81.20	77.80	73.70	93.90	81.70
SagNet [38]	83.58	77.66	76.30	95.47	83.25
DAML [49]	83.00	78.10	74.10	95.60	82.70
StableNet [69]	81.74	79.91	80.50	96.53	84.69
W2D [18]	-	-	-	-	83.4
BatchFormer [17]	84.8	75.3	81.1	93.6	83.7
ITL-Net [13]	83.9	78.9	94.8	80.1	84.4
Style Neophile [21]	84.41	79.25	94.93	83.27	85.47
RSC [19]	80.73	79.22	81.48	94.16	83.90
Ours	83.64	80.03	84.37	95.32	85.84

Table 2: Classification accuracy (%) on the Office-Home dataset. The bold numbers indicate the best performance averaged across all domains, second best is underlined.

Office-Home	Art	Clipart	Product	Real	Avg \uparrow
baseline [3]	52.15	45.86	70.86	73.15	60.51
Jigen [3]	53.04	47.51	71.47	72.79	61.20
DSON [45]	59.37	45.70	71.84	74.68	62.90
L2A-OT [71]	60.60	50.10	74.80	77.00	<u>65.60</u>
CCSA [36]	59.9	49.9	74.1	75.7	64.9
MMD-AAE [27]	56.5	47.3	72.1	74.8	62.7
CrossGrad [46]	58.4	49.4	73.9	75.8	64.4
SagNet [38]	60.20	45.38	70.42	73.38	62.34
W2D [18]	-	-	-	-	63.5
BatchFormer [17]	57.8	51.0	73.4	75.1	64.3
Style Neophile [21]	59.55	55.01	73.57	75.52	65.89
RSC [19]	58.42	47.90	71.63	74.54	63.12
Ours	59.46	52.81	73.85	74.98	65.28

total of over half a million images is grouped into 345 categories. To facilitate faster prototyping and experimentation, miniDomainNet [73] has been introduced which reduces DomainNet to 4 domains: *clipart*, *painting*, *real*, *sketch*, with 126 classes and nearly a quarter million images. We divided the images into training and testing sets following the original setup [73]. The complexity of miniDomainNet significantly surpasses both PACS and Office-Home, thus making experimental validation much more meaningful.

As showed in Table 3, our method yet again outperforms RSC and other state-of-the-art models, establishing itself as a strong competitor in domain generalization setting. In the supplementary material, we present more experimental results showcasing the versatility of our approach. Additionally, we provide further insights into the assumption that

Table 3: Classification accuracy (%) on the miniDomainNet dataset. The bold numbers indicate the best performance averaged across all domains, second best is underlined.

miniDomainNet	Clipart	Painting	Real	Sketch	Avg \uparrow
DANN [12]	65.55	46.27	58.68	47.88	54.60
DCTN [62]	62.06	48.79	58.85	48.25	54.49
MCD [43]	62.91	45.77	57.57	45.88	53.03
MME [42]	68.09	47.14	63.33	43.50	55.52
DAEL [73]	69.95	55.13	66.11	55.72	61.73
CMSDA [44]	71.38	53.76	66.23	56.24	61.90
FAUST [25]	68.1	52.2	68.7	59.1	62.0
RSC [19]	65.34	59.72	66.70	58.94	<u>62.67</u>
Ours	69.41	61.48	66.81	62.83	65.13

iterative removal of the most predictive parameters leads to the learning less dominant features.

4.5. OCIM Face Anti-Spoofing

Finally, to further demonstrate the applicability of our model, we tackle a different computer vision task of face anti-spoofing. It is a binary classification problem where the goal is to distinguish between real and spoofed faces. We validate our approach on a popular OCIM benchmark consisting of 4 different datasets: OULU-NPU [2], CASIA-FASD [70], MSU-MFSD [6], and REPLAY-ATTACK [61]. Various acquisition and presentation attack devices along with multiple lightning conditions contribute to the domain shift present in this benchmark. Each dataset provides videos of approximately 50 unique subjects which are then processed into frames. We follow the OCIM protocols proposed in [47] for cross-dataset testing.

Here, we base our model on Shuffled Style Assembly Network (SSAN) [59]. Concretely, we introduce a GSNR-guided dropout procedure into the feature generator and leave the remaining of SSAN [59] unchanged. Similarly, we augment the EPCR [60] model. Results presented in Table 4 shows that our modification improves the results according to both metrics and on nearly all protocols.

4.6. Comparison with RSC

To facilitate a fair comparison between RSC and our approach, we combine classical regularization method such as DropOut [51], DropBlock [14] and SpatialDropout [52] with different methods of dropout mask construction: based on random seeds, as done in the original approach, based on magnitude of the gradients, as done in RSC, and based on gradient-signal-to-noise ratio of the network’s parameters. Ablation study presented in Table 5 shows that dropout mask construction guided by GSNR of networks parameters outperforms other approaches by a significant margin.

4.7. Model Analysis

Ablation Study: We compare various dropout strategies on PACS dataset, shown in Table 6. Clearly, methods that mute activations based on *most important* features outperform those relying on random selection. As reported in [74], classical regularization methods fail to discover new patterns when domain shift has been introduced to the data. Applying dropout to multiple ResNet blocks further improves the results, while best performance is obtained when dropout ratio is learned. We demonstrate the impact of intermediate masks $M^{(1)}$ and $M^{(2)}$ of our approach in the supplementary material.

GSNR of Network Parameters over Time: To validate whether our approach enhances the overall GSNR of network’s parameters over time, we register an average GSNR of all parameters over the entire training period. Figure 2 shows that the model equipped with our procedure has higher gradient-signal-to-noise ratio compared to the baseline, confirming that our approach improves models’ GSNR. GSNR evaluates the generalization ability during training, and monitors the intermediate training process. At full convergence, the difference of signal and noise across two training models becomes negligible as the *noise* is less likely to occur, reaching a similar magnitude for both models. However, the paths of how the two models are updated is quite different, resulting in different levels of generalization. Both GSNR and baseline were trained for an extended number of epochs to highlight the dynamics of overfitting.

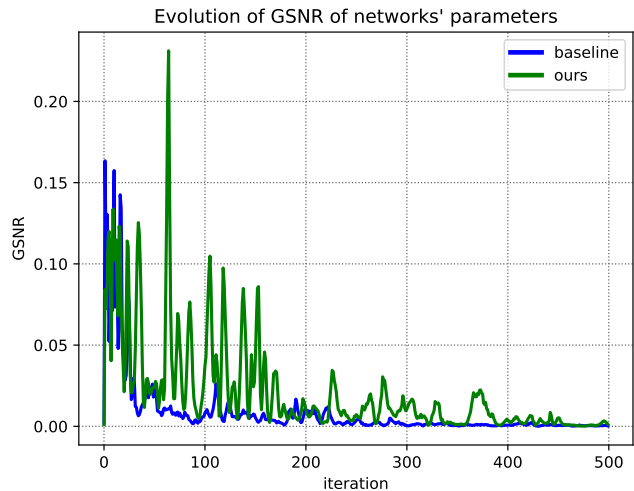


Figure 2: Our approach enhances networks gradient-signal-to-noise ratio over time, compared to the baseline.

Learned dropout ratios: In this section, we analyze the dropout ratios learned by our model and those found by the grid search approach. Figure 3 depicts an accuracy distribu-

Table 4: Classification accuracy (%) on the OCIM dataset. The bold numbers indicate the best average performance measured by two different metrics: HTER and AUC.

Method	O&C&I to M		O&M&I to C		O&C&M to I		I&C&M to O		Avg HTER ↓	Avg AUC ↑
	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑	HTER(%) ↓	AUC(%) ↑		
MMD-AAE [27]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08	36.05	69.93
MADDG [47]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02	23.09	84.39
DR-MD-Net [56]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47	20.64	86.43
RFMeta [48]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16	16.97	90.94
D2AM [5]	12.70	95.66	20.98	85.58	15.43	91.22	15.27	90.87	16.09	90.83
SDA [57]	15.40	91.80	24.50	84.40	15.60	90.10	23.10	84.30	19.65	87.64
DRDG [34]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75	15.66	92.03
ANRL [33]	10.83	96.75	17.83	89.26	16.03	91.04	15.67	91.90	15.09	92.23
SSDG [20]	9.63	92.82	13.38	90.97	15.15	91.89	19.89	86.29	14.51	90.49
SSAN [59]	2.50	97.50	13.59	93.67	17.62	83.91	19.15	87.95	13.21	90.75
SSAN + Ours	2.50	95.50	10.94	95.17	12.25	93.60	18.42	88.63	11.02	93.22
EPCR [60]	7.50	93.73	9.89	93.61	11.93	91.96	16.12	88.62	11.36	91.98
EPCR + Ours	7.02	94.52	9.32	95.16	10.34	94.39	15.21	89.05	10.47	93.28

Table 5: Ablation study: Classification accuracy (%) on PACS dataset with various approaches to dropout mask construction: guided by Bernoulli samples *rand*, magnitude of gradients of network’s parameters *mag*, and their gradient-signal-to-noise ratios *GSNR*

PACS	artpaint	cartoon	sketch	photo	Avg ↑
Dropout w/ rand [51]	76.93	75.60	73.66	93.35	79.88
Dropout w/ mag	80.12	76.49	80.75	94.67	83.01
Dropout w/ GSNR	81.44	78.02	82.33	94.07	83.96
DropBlock w/ rand [14]	76.4	75.4	69.0	95.9	79.2
DropBlock w/ mag	79.34	79.01	79.20	95.08	83.15
DropBlock w/ GSNR	81.94	80.07	80.78	94.95	84.43
SpatialDropout w/ rand [52]	73.82	78.07	77.49	93.25	80.65
SpatialDropout w/ mag	79.73	78.37	79.81	94.97	83.22
SpatialDropout w/ GSNR	79.68	78.79	79.47	95.32	83.31
RSC-rand	76.80	77.71	78.11	94.31	81.73
RSC-mag	80.73	79.22	81.48	94.16	83.90
RSC-GSNR	81.15	80.16	81.67	94.55	84.38

Table 6: Ablation study: Classification accuracy (%) on PACS dataset with various dropout strategies.

PACS	artpaint	cartoon	sketch	photo	Avg ↑
Baseline	77.0	75.9	69.2	96.0	79.5
CutOut [8]	74.9	74.9	67.7	95.9	78.3
MixUp [67]	76.8	74.9	95.8	66.6	78.5
CutMix [65]	74.6	71.8	95.6	65.3	76.8
Manifold Mixup [55]	75.6	70.1	65.4	93.5	76.2
DropBlock (DB) [14]	76.4	75.4	69.0	95.9	79.2
DB + GSNR	81.94	80.07	80.78	94.95	84.43
DB + GSNR + grid search	83.34	80.33	82.56	95.26	85.37
DB + GSNR + meta	83.64	80.03	84.37	95.32	85.84

tion gathered by 6^4 different configurations while Figures 4 and 5 show the evolution of dropout ratio p for our meta-learning based approach. It’s shown that different ResNet blocks and different domains favour different dropout ratios supporting our claim that p should be learned. Note that the brute force approach is infeasible as it requires to train models with countless variations of p . It should rather be viewed as an oracle approach that motivates leveraging the learning-to-learn techniques for alleviating the compu-

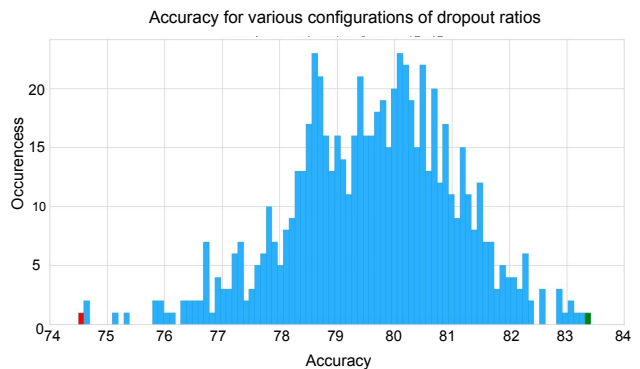


Figure 3: Accuracy distribution for different dropout ratio configurations. Worst configuration, obtained by (0, 0.4, 0.3, 0) is marked in red, while best, obtained by (0, 0, 0.2, 0.4) is marked in green.

tational burden of finding the best configuration of p . Finally, setting the dropout ratio to a fixed value can be far from optimal as shown in Figure 3.

Stiffness: We investigate the generalization capabilities of our model through the lenses of stiffness [11], *i.e.* by inspecting how much a small gradient update on one data point affects the loss on another. Specifically, the better the pairs of gradients are aligned, the higher the stiffness of the network which is linked to better generalization. We compute two variants of stiffness: intra-class and inter-class. For intra-class stiffness, where data points x_i and x_j share the same label $y_i = y_j$, we measure the alignment of the corresponding gradients g_i and g_j using the *cosine* formula:

$$S_{\cos}((x_i, y_i), (x_j, y_j); f) = \mathbb{E}[\frac{g_i}{\|g_i\|} \cdot \frac{g_j}{\|g_j\|}]. \quad (23)$$

To compute inter-class stiffness, where labels y_i and y_j are different, we use the *sign* formula:

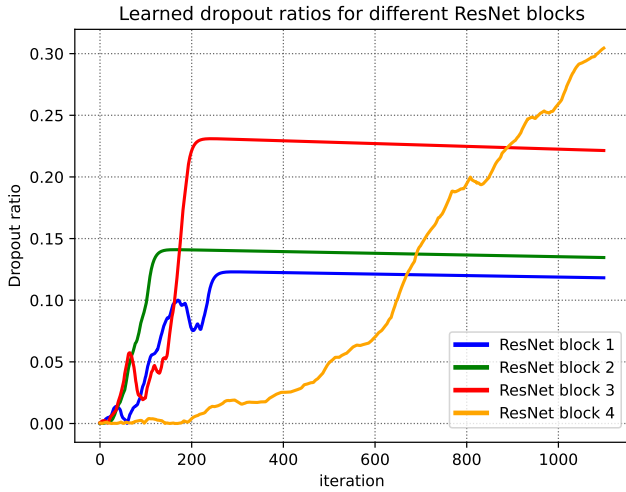


Figure 4: Evolution of dropout ratio over time for different ResNet blocks.

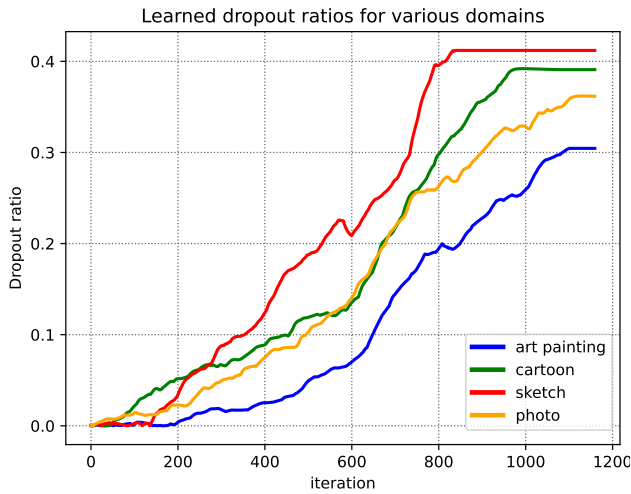


Figure 5: Evolution of dropout ratio over time for ResNet block 4 for different domains.

$$S_{\text{sign}}((x_i, y_i), (x_j, y_j); f) = \mathbb{E}[\text{sign}(g_i, g_j)]. \quad (24)$$

We show that our model exhibits improved inter-class and intra-class stiffness in Figures 7 and 6, respectively.

5. Conclusion and Future Work

In this work, we introduce Meta-GSNR: a domain generalization method guided by gradient-signal-to-noise ratio of network’s parameters. This DropBlock-based regularization procedure reduces the generalization gap by iteratively muting activations of parameters with highest GSNR values. We alleviate manual selection of dropout ratios by

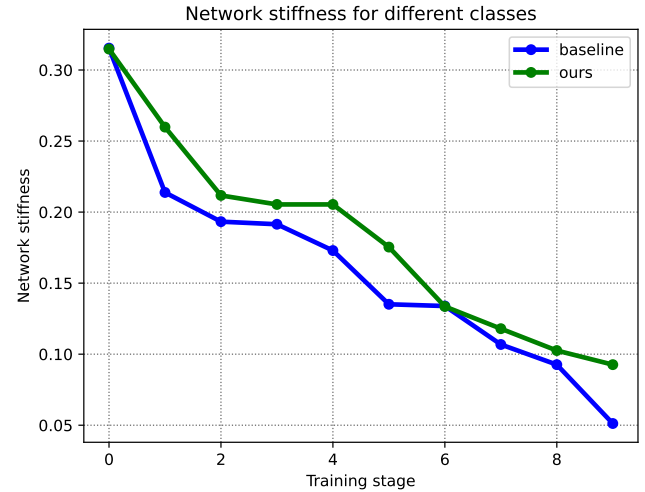


Figure 6: Evolution of *inter-class* stiffness measured by Equation 24. Here, we only consider pairs with different labels.

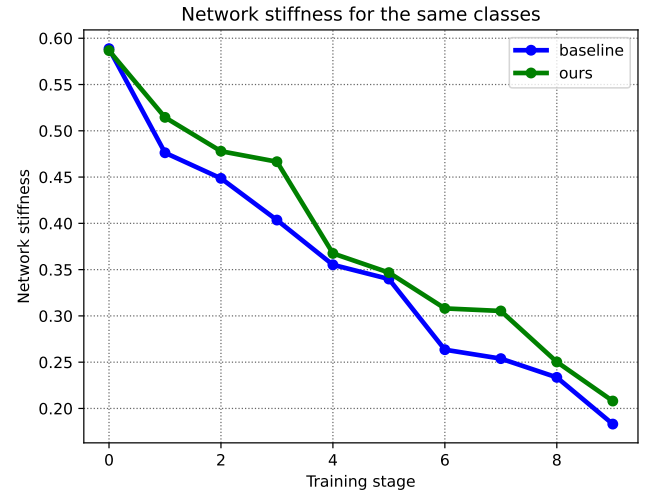


Figure 7: Evolution of *intra-class* stiffness measured by Equation 23. Here, we only consider pairs with the same labels.

leveraging a learning-to-learn technique. Extensive experiments conducted on standard classification and face recognition benchmark datasets of PACS, Office-Home, miniDomainNet, and OCIM demonstrate the effectiveness of our approach.

In future, we would like to investigate other gradient-based regularization metrics such as stiffness, and their effect on model generalizability.

Acknowledgements. This work was a part of Mateusz Michalkiewicz’s internship at NEC Labs America.

References

- [1] Haoyue Bai, Fengwei Zhou, Lanqing Hong, Nanyang Ye, S-H Gary Chan, and Zhenguo Li. Nas-ood: Neural architecture search for out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8320–8329, 2021. [6](#)
- [2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017. [2, 7](#)
- [3] 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. [6](#)
- [4] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pages 301–318. Springer, 2020. [6](#)
- [5] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 of 2, pages 1132–1139, 2021. [8](#)
- [6] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. [2, 7](#)
- [7] 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. Ieee, 2009. [5](#)
- [8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. [2, 8](#)
- [9] Masoud Faraki, Xiang Yu, Yi-Hsuan Tsai, Yumin Suh, and Manmohan Chandraker. Cross-domain similarity learning for face recognition in unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15292–15301, 2021. [2](#)
- [10] 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](#)
- [11] Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Sridhar Narayanan. Stiffness: A new perspective on generalization in neural networks. *arXiv preprint arXiv:1901.09491*, 2019. [3, 8](#)
- [12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. [7](#)
- [13] Boyan Gao, Henry Gouk, Yongxin Yang, and Timothy Hospedales. Loss function learning for domain generalization by implicit gradient. In *International Conference on Machine Learning*, pages 7002–7016. PMLR, 2022. [6](#)
- [14] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018. [1, 2, 5, 7, 8](#)
- [15] Jintao Guo, Lei Qi, Yinghuan Shi, and Yang Gao. Domain generalization via progressive layer-wise and channel-wise dropout. *arXiv preprint arXiv:2112.03676*, 2021. [6](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [3, 5, 6](#)
- [17] Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7256–7266, 2022. [6](#)
- [18] Zeyi Huang, Haohan Wang, Dong Huang, Yong Jae Lee, and Eric P Xing. The two dimensions of worst-case training and their integrated effect for out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9641, 2022. [6](#)
- [19] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. [1, 2, 3, 5, 6, 7](#)
- [20] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. [8](#)
- [21] 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*, pages 7130–7140, 2022. [6](#)
- [22] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017. [2](#)
- [23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. [6](#)
- [24] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016. [2](#)
- [25] JoonHo Lee and Gyemin Lee. Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation. *Neural Networks*, 161:682–692, 2023. [7](#)
- [26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. [2, 6](#)

- [27] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 6, 8
- [28] Jingling Li, Yanchao Sun, Jiahao Su, Taiji Suzuki, and Furong Huang. Understanding generalization in deep learning via tensor methods. In *International Conference on Artificial Intelligence and Statistics*, pages 504–515. PMLR, 2020. 3
- [29] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021. 6
- [30] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017. 2, 4
- [31] Chang Liu, Xiang Yu, Yi-Hsuan Tsai, Masoud Faraki, Ramin Moslemi, Manmohan Chandraker, and Yun Fu. Learning to learn across diverse data biases in deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4072–4082, 2022. 2, 4
- [32] Jinlong Liu, Guoqing Jiang, Yunzhi Bai, Ting Chen, and Huayan Wang. Understanding why neural networks generalize well through gsnr of parameters. *arXiv preprint arXiv:2001.07384*, 2020. 1, 3
- [33] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1469–1477, 2021. 8
- [34] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. In *IJCAI*, 2021. 8
- [35] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 of 07, pages 11749–11756, 2020. 6
- [36] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017. 6
- [37] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 2
- [38] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 6
- [39] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2
- [40] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 2, 6
- [41] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 3
- [42] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 7
- [43] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 7
- [44] Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization. *arXiv preprint arXiv:2106.16093*, 2021. 7
- [45] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 68–83. Springer, 2020. 6
- [46] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018. 6
- [47] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10031, 2019. 7, 8
- [48] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, volume 34 of 07, pages 11974–11981, 2020. 8
- [49] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021. 6
- [50] Christian Simon, Masoud Faraki, Yi-Hsuan Tsai, Xiang Yu, Samuel Schulter, Yumin Suh, Mehrtash Harandi, and Manmohan Chandraker. On generalizing beyond domains in cross-domain continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9265–9274, 2022. 1
- [51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way

- to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 1, 2, 7, 8
- [52] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015. 2, 7, 8
- [53] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 2
- [54] Hemant Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2, 6
- [55] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019. 8
- [56] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020. 8
- [57] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 4 in 35, pages 2746–2754, 2021. 8
- [58] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021. 6
- [59] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4123–4133, 2022. 6, 7, 8
- [60] Zezheng Wang, Zitong Yu, Xun Wang, Yunxiao Qin, Jiahong Li, Chenxu Zhao, Xin Liu, and Zhen Lei. Consistency regularization for deep face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:1127–1140, 2023. 6, 7, 8
- [61] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 2, 7
- [62] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3964–3973, 2018. 7
- [63] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 2
- [64] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020. 2
- [65] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 8
- [66] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 3
- [67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 8
- [68] Marvin Zhang, Henrik Marklund, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. *CoRR*, abs/2007.02931, 2020. 2
- [69] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyang Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021. 6
- [70] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 2, 7
- [71] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, pages 561–578. Springer, 2020. 6
- [72] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020. 3, 6
- [73] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021. 6, 7
- [74] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 1, 2, 7
- [75] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 2