# Sharpness-Aware Model-Agnostic Long-Tailed Domain Generalization

**Houcheng Su[1*], Weihao Luo[2*], Daixian Liu [3], Mengzhu Wang[4†], Jing Tang[4], Junyang Chen[5], Cong Wang[6], Zhenghan Chen[7]**

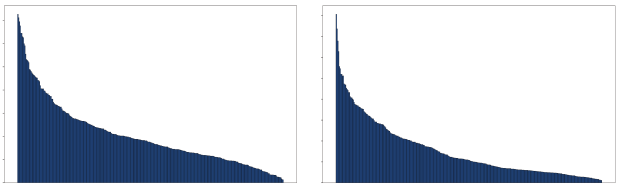
[1]University of Macau
[2]Donghua University
[3]Sichuan Agricultural University
[4]Hebei University of Technology
[5]Shenzhen University
[6]The Hong Kong Polytechnic University
[7]Peking University
{mc25695,yb77403}@umac.mo, luowh@mail.dhu.edu.cn, 202105787@stu.sicau.edu.cn, wangmengzhu.wmz@alibaba-inc.com, focusers@163.com, supercong.wang@connect.polyu.hk, pandaarych@gmail.com

## Abstract

Domain Generalization (DG) aims to improve the generalization ability of models trained on a specific group of source domains, enabling them to perform well on new, unseen target domains. Recent studies have shown that methods that converge to smooth optima can enhance the generalization performance of supervised learning tasks such as classification. In this study, we examine the impact of smoothness-enhancing formulations on domain adversarial training, which combines task loss and adversarial loss objectives. Our approach leverages the fact that converging to a smooth minimum with respect to task loss can stabilize the task loss and lead to better performance on unseen domains. Furthermore, we recognize that the distribution of objects in the real world often follows a long-tailed class distribution, resulting in a mismatch between machine learning models and our expectations of their performance on all classes of datasets with long-tailed class distributions. To address this issue, we consider the domain generalization problem from the perspective of the long-tail distribution and propose using the maximum square loss to balance different classes which can improve model generalizability. Our method's effectiveness is demonstrated through comparisons with state-of-the-art methods on various domain generalization datasets. Code: https://github.com/bamboosir920/SAMALTDG.

## Introduction

Deep learning approaches have proven to be highly effective in computer vision tasks, especially when the source and target data are independently and identically distributed. However, these methods often suffer from reduced performance when applied to new target domains. To address this, domain generalization (DG) (Zhang et al. 2022a; Qiao, Zhao, and Peng 2020; Balaji, Sankaranarayanan, and Chellappa 2018) techniques aim to train models using source data that can perform well on new domains without retraining. Numerous DG methods have been developed over the past decade, including those based on domain alignment (Muandet, Balduzzi, and Schölkopf 2013), meta-training (Li et al. 2018a), and data augmentation (Wang et al. 2022b). Despite the many approaches that have been proposed, a recent study called DomainBed (Gulrajani and Lopez-Paz 2020) found that the naive DG method via entropy regularization (DG_via_ER) (Zhao et al. 2020) can perform better than most other DG methods under fair evaluation conditions. Nonetheless, simply minimizing empirical loss on a non-convex loss landscape is typically insufficient to achieve robust generalization. As such, DG_via_ER may overfit to the training data and converge to sharp local minima.



(a)Infograph (b)Sketch

Figure 1: Data distributions of two benchmark datasets. (a) shows the number of different classes in "Infograph", (b) shows the number of different classes in "Sketch".

Various recent studies, such as sharpness-aware minimization (SAM) (Foret et al. 2020), aims to improve the model's performance by minimizing the sharpness measure of the loss landscape. The loss function to be minimized, $\mathcal{L}_\theta$, depends on the neural network's parameters $\theta$ (e.g., cross-entropy loss for classification). SAM computes an adversarial weight perturbation $\epsilon$ to maximize the empirical risk $\mathcal{L}_\theta$, followed by minimizing the loss of the perturbed network. SAM's objective is to minimize the maximum loss around the model parameter $\theta$. Since the min-max optimization problem is highly complex, SAM approximates $\mathcal{L}_\theta$ instead. Inspired by SAM, we aim to improve the model's generalization ability by minimizing sharpness. However, recent

*These authors contributed equally.
†This is the corresponding author

analyses (Rangwani et al. 2022b) have revealed that SAM fails to prevent tail classes from converging to saddle points in high curvature regions, resulting in poor generalization.

We observed that most existing datasets exhibit a long-tailed distribution, yet domain generalization methods seldom consider this perspective. As demonstrated in Figure. 1, we investigated the "Infograph" and "Sketch" domains from the large-scale DomainNet dataset and found that both displayed pronounced long-tailed distributions. Recently, entropy minimization techniques in semi-supervised learning (Grandvalet and Bengio 2004; Chen, Xue, and Cai 2019), which encourage clear cluster assignments, have become popular. Upon analyzing the gradient of the entropy minimization method in domain generalization (Chen, Xue, and Cai 2019), we discovered that higher prediction probabilities induce larger gradients for target samples. Adopting the assumption from self-training that target samples with higher prediction probabilities are more accurate leads to areas with high accuracy receiving sufficient training, while areas with low accuracy do not. Consequently, the entropy minimization method enables adequate training of samples that are easy to transfer, which obstructs the training of samples that are difficult to transfer. This issue in entropy minimization is known as probability imbalance: classes that are easy to transfer have higher probabilities, resulting in much larger gradients than classes that are difficult to transfer.

In this paper, we introduce a new loss, the maximum squares loss (Chen, Xue, and Cai 2019), to tackle the probability imbalance problem. Since the loss of the maximum square has a linearly increasing gradient, it can prevent high-confident areas from producing excessive gradients. We leverage the popular method DG_via_ER (Zhao et al. 2020) and minimize the sharpness measure of the classification loss as our baseline. We also demonstrate the effectiveness of our approach by conducting comprehensive experiments on several benchmarks.

## Related Work

**Domain Generalization:** Domain generalization (DG) aims to transfer the learning task from multiple source domains and generalize to unseen target domains (Zhou et al. 2021). Early research in this field concentrated on the concept of distribution alignment, akin to domain adaptation, utilizing kernel methods (Muandet, Balduzzi, and Schölkopf 2013; Ghifary et al. 2016) and domain-adversarial learning (Li et al. 2018c,b) to tackle the issue. Later investigations shifted the focus towards the extraction of domain-invariant features across multiple source domains to establish domain invariance (Wang et al. 2022b, 2021a, 2022a, 2023). A number of strategies have employed meta-learning for the derivation of regularization strategies to address the DG problem (Li et al. 2018a; Balaji, Sankaranarayanan, and Chellappa 2018). Yao et al. found the direct application of contrastive-based methods, though used to resolve domain generalization, could prove ineffective (Yao et al. 2022), suggesting the substitution of original sample-to-sample relations with proxy-to-sample relations.

A myriad of techniques make up the recent advancements in domain generalization. Yang et al. put forth the proposal of Adversarial Teacher-Student Representation Learning to create domain-generalizable representations by exploring and generating out-of-source data distributions (Yang et al. 2021). Xu et al. hypothesized that Fourier phase information, which encompasses high-level semantics, is resistant to domain shifts, leading to the introduction of a novel Fourier-based data augmentation strategy (Xu et al. 2021). Zhao et al. employed an entropy regularization term to calculate the dependency between class labels and learned features (Zhao et al. 2020). Zhang et al. suggested that domain generalization could be resolved by matching exact feature distributions (Zhang et al. 2022b). Wang et al. adopted a multi-task learning paradigm to learn feature embedding that generalizes across domains simultaneously from extrinsic relationship supervision and intrinsic self-supervision for images from multi-source domains (Wang et al. 2020). Zhang et al. offered a method to quantify and enhance transferability with an efficient algorithm for the learning of transferable features (Zhang et al. 2021).

Recent studies in DG have expanded into the area of Single Domain Generalization, which concentrates on generalization from a lone source domain to unseen target domains (Qiao, Zhao, and Peng 2020; Wan et al. 2022). LDMI (Wang et al. 2021b) propose a style-complement module to enhance the generalization power of the model by synthesizing images from diverse distributions that are complementary to the source ones. TASD (Liu et al. 2022) present a novel approach to address the challenging single domain generalization problem for medical image segmentation, by explicitly exploiting the general semantic shape priors that are extractable from single-domain data and are generalizable across domains to assist domain generalization under the worst-case scenario. This particular line of research has shown promise in utilizing a single domain to achieve effective generalization, a factor that is particularly relevant when faced with limited data or the unavailability of multiple source domains.

Recently, the task of Multi-Domain Long-Tailed Recognition (MDLT) was formalized by Yang et al. (Yang, Wang, and Katabi 2022). MDLT tackles the challenges associated with label imbalance, domain shift, and varying label distributions across domains. By generalizing across all domain-class pairs, MDLT provides a more comprehensive solution for real-world recognition problems that involve multiple domains and long-tailed distributions. Similar to these methods, We are considering addressing the domain generalization problem from the perspective of long-tailed distribution.

## Method

Assume $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label spaces, respectively. In domain generalization, the subject encompasses $K$ source domains $\{\mathcal{D}_i\}_{i=1}^{K}$ and $L$ target domains $\{\mathcal{D}_i\}_{i=K+1}^{L+K}$, and the objective is to generalize the model trained on source domain data to unseen target domains. Here, $P_i(X, Y)$ denotes the joint distribution of the $i$th domain. During training, there exist $K$ datasets $\{S_i\}_{i=1}^{K}$ with $N_i$ samples from the $i^{th}$ domain. In the testing stage, the model's generalization capabilities are assessed on $L$

datasets sampled from the $L$ target domains. This paper specifically focuses on image classification domain generalization, where $\mathcal{Y}$ comprises $C$ discrete labels $\{1, 2, \cdots, C\}$.

## Domain Generalization via Entropy Regularization (DG_via_ER)

In this paper, we introduce the utilization of the domain generalization method in DG_via_ER (Zhao et al. 2020). Regarding the classification topic, our model comprises of a feature extractor denoted as $F$ with parameters $\theta$, and a classifier called $T$ with parameters $\phi$. We can achieve optimal values of $\theta$ and $\phi$ by minimizing the cross-entropy loss function over $K$ source datasets.

$$\begin{aligned}\min_{F,T} \mathcal{L}_{cls}(\theta, \phi) &= -\sum_{i=1}^{K} \mathop{\mathbb{E}}_{(X,Y)\sim P_i(X,Y)} \left[\log\left(Q^T(Y \mid F(X))\right)\right] \\ &= -\sum_{i=1}^{K}\sum_{j=1}^{N_i} \mathbf{y}_j^{(i)} \cdot \log\left(T\left(F\left(\mathbf{x}_j^{(i)}\right)\right)\right),\end{aligned} \tag{1}$$

Here, $\mathbf{y}_j^{(i)}$ refers to the one-hot vector representation of the class label $y_j^{(i)}$. The symbol "$\cdot$" represents the dot product operation, while $Q^T(Y \mid F(X))$ indicates the predicted label distribution that corresponds to the given domain $i$.

Despite being optimized solely with the classification loss, the model is incapable of acquiring domain-invariant features, leading to challenges in generalizing to unfamiliar domains. However, utilizing adversarial learning can help mitigate this problem. This involves introducing a domain discriminator parameterized by $\psi$, and training it and $F$ in a minimax game as follows:

$$\begin{aligned}\min_{F}\max_{D} \mathcal{L}_{adv}(\theta, \psi) &= \sum_{i=1}^{K} \mathop{\mathbb{E}}_{X\sim P_i(X)} [\log D(F(X))] \\ &= \sum_{i=1}^{K}\sum_{j=1}^{N_i} \mathbf{d}_j^{(i)} \cdot \log\left(D\left(F\left(\mathbf{x}_j^{(i)}\right)\right)\right)\end{aligned} \tag{2}$$

Here, $\mathbf{d}_j^{(i)}$ denotes the one-hot encoding of the domain labels $i$.

While optimizing Eq. 2 may result in invariant marginal distributions $P_1(F(X)) = P_2(F(X)) = \cdots = P_K(F(X))$, it does not ensure that the conditional distribution $P(Y|F(X))$ remains invariant across domains. As a result, the model's ability to generalize may suffer. To address this issue, DG_via_ER (Zhao et al. 2020) proposes the use of entropy regularization for domain generalization.

To regularize the feature distributions, DG_via_ER (Zhao et al. 2020) proposes minimizing the KL divergence between the conditional distribution $P_i(Y \mid F(X))$ in the $i^{th}$ domain and the conditional distribution $Q^T(Y \mid X)$. Here, $P_i(Y \mid F(X))$ refers to the predicted label distribution based on the learned features. By aligning any conditional distribution $P_i(Y \mid F(X))$ with a common distribution $Q^T(Y \mid F(X))$, DG_via_ER can obtain a domain-invariant conditional distribution $P(Y midF(X))$. The optimization problem is as follows:

$$\mathcal{L}_{er} = \min_{F,T} \sum_{i=1}^{K} KL\left(P_i(Y \mid F(X)) \| Q^T(Y \mid F(X))\right) \tag{3}$$

Although DG_via_ER (Foret et al. 2020) can learn domain-invariant features from the perspective of adversarial learning, it ignores the search for optimal extremal points, which may impair its generalization ability. Inspired by the recent popular model SAM, we consider seeking a region with low loss values by adding a small perturbation to the models which can further improve the generalization of the model.

## Smoothing Loss Landscape

This section introduces the losses based on Sharpness Aware Minimization (SAM) (Rangwani et al. 2022a). SAM aims to find a smoother minimum by utilizing the following objective, which is presented formally below:

$$\min_{\theta} \max_{||\epsilon|| \le \rho} \mathcal{L}_{obj}(\theta + \epsilon) \tag{4}$$

Here, $\mathcal{L}_{obj}$ denotes the objective function that needs to be minimized, and $\rho \ge 0$ is a hyperparameter that defines the maximum norm for $\epsilon$. As finding the exact solution for the inner maximization is challenging, SAM maximizes the first-order approximation:

$$\begin{aligned}\hat{\epsilon}(\theta) &\approx \arg\max_{||\epsilon|| \le \rho} L_{obj}(\theta) + \epsilon^{\mathbf{T}} \nabla_\theta L_{obj}(\theta) \\ &= \rho \nabla_\theta L_{obj}(\theta) / ||\nabla_\theta L_{obj}(\theta)||_2\end{aligned} \tag{5}$$

The $\hat{\epsilon}(\theta)$ is added to the weights $\theta$. The gradient update for $\theta$ is then computed as $\nabla_\theta \mathcal{L}_{obj}(\theta)|_{\theta+\hat{\epsilon}(\theta)}$. The above procedure can be seen as a generic smoothness-enhancing formulation for any $\mathcal{L}_{obj}$. We now analogously introduce the sharpness-aware source risk for finding a smooth minima:

$$\max_{||\epsilon|| \le \rho} R_S^l(h_{\theta+\epsilon}) = \max_{||\epsilon|| \le \rho} \mathbb{E}_{x\sim P_S}[\, l(h_{\theta+\epsilon}(x), f(x))] \tag{6}$$

We also now define the sharpness aware discrepancy estimation objective below:

$$\max_{\Phi} \min_{||\epsilon|| \le \rho} d_S^{\Phi+\epsilon} \tag{7}$$

As $d_S^\Phi$ is to be maximized the sharpness aware objective will have $\min_{||\epsilon|| \le \rho}$ instead of $\max_{||\epsilon|| \le \rho}$, as it needs to find smoother maxima. We now theoretically analyze the difference in discrepancy estimation for smooth version $d_S^{\Phi''}$ (Eq. 7) in comparison to non-smooth version $d_S^{\Phi'}$. Assuming $\mathcal{D}_\Phi$ is a $L$-smooth function (common assumption for non-convex optimization), $\eta$ is a small constant and $d_S^*$ the optimal discrepancy, the theorem states.

## Maximum Square Loss

To learn more diverse features, we try to leverage the Shannon entropy of the target sample prediction. Thus, the objective function for the source sample is :

$$\mathcal{L}_S(x_s) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} p_s^{n,c} \log\left(p_s^{n,c}\right) \tag{8}$$

We draw inspiration from MaxSquare (Chen, Xue, and Cai 2019) and utilize maximum squares loss to prevent the training process from being dominated by easily transferable samples. To simplify matters, we focus on the binary classification scenario and present the corresponding entropy formula and gradient function.

$$\begin{aligned} H\left(p \mid x_s\right) &= -p \log p - (1-p) \log(1-p), \\ \left|\frac{dH}{dp}\right| &= |\log p - \log(1-p)|. \end{aligned} \tag{9}$$

The gradient obtained from Eq. 9 is significantly larger for high-probability points than for intermediate points. This is the fundamental principle behind the entropy minimization method, which guides the training of target samples based on the assumption that the high probability area is more accurate. To achieve a uniform probability distribution, we therefore consider the use of maximum square loss.

$$\mathcal{L}_m\left(x_s\right) = -\frac{1}{2N} \sum_{n=1}^{N} \sum_{c=1}^{C} \left(p_s^{n,c}\right)^2 \tag{10}$$

In the scenario of binary classification, we can express the maximum squares loss and its corresponding gradient function as follows:

$$\begin{aligned} MS\left(p \mid x_s\right) &= -p^2 - (1-p)^2, \\ \left|\frac{dMS}{dp}\right| &= |4p - 2|. \end{aligned} \tag{11}$$

As indicated by the above equation, the gradient of the maximum squares loss increases linearly, resulting in a more balanced gradient for different classes compared to the entropy minimization method in the target domain. While areas with higher confidence still possess larger gradients, their dominant effects are reduced, allowing other difficult classes to obtain training gradients as well. By utilizing the maximum squares loss, we can alleviate the probability imbalance present in entropy minimization.

### Overall Formulation

Combining all the loss functions together, we can get our full objective as:

$$\mathcal{L}_{ours} = \mathcal{L}_{adv} + \mathcal{L}_{er} + \gamma \mathcal{L}_m \tag{12}$$

where $\gamma$ controls the trade-off between the classification loss and maximum square loss.

## Experiment

In this section, we investigate the effectiveness of our proposed improvements on three state-of-the-art methods, to demonstrate the validity of our approach. Comparative experiments are conducted across four datasets: PACS (Li et al. 2017), OfficeHome (Venkateswara et al. 2017), DigitDG (Zhou et al. 2020), and DomainNet (Peng et al. 2019). In addition, we perform ablation studies to facilitate a thorough discourse on our methodology.

### Datasets and Settings

**PACS:** PACS (Li et al. 2017) is proposed specially for domain generalization. It contains four domains, i.e., Photo (P), Art Painting (A), Cartoon (C), and Sketch (S), and seven categories: dog, elephant, giraffe, guitar, house, horse, and person. We use the same training and validation split as presented in (Li et al. 2017) for a fair comparison. We randomly split each domain into 90% for training and 10% for validation.

**OfficeHome:** OfficeHome (Venkateswara et al. 2017) is an object recognition benchmark including 15,500 images of 65 classes from four domains (Art, Clipart, Product, Real-World). The domain shift mainly comes from image styles and viewpoints but is much smaller than PACS. Following (Carlucci et al. 2019), we randomly split each domain into 90% for training and 10% for validation.

**DigitsDG:** DigitsDG (Zhou et al. 2020) is a digit recognition benchmark consisting of four classical datasets MNIST (Carlucci et al. 2019), MNIST-M (Ganin and Lempitsky 2015), SVHN (Netzer et al. 2011), SYN (Ganin and Lempitsky 2015). The four datasets mainly differ in font style, background and image quality. We use the original train validation split in (Zhou et al. 2020) with 600 images per class per dataset. We randomly split each domain into 90% for training and 10% for validation.

**DomainNet:** DomainNet (Peng et al. 2019) is a dataset of common objects in six different domains. All domains include 345 categories (classes) of objects such as Bracelet, plane, bird and cello. The domains include **Clipart:** collection of Clipart images; **Real:** photos and real-world images; **sketch:** sketches of specific objects; **Infograph:** infographic images with a specific object; **Painting:** painting artistic depictions of objects in the form of paintings and **Quickdraw:** drawings of the worldwide players of the game. For data sets, we adopted the default partitioning method of data sets, with 80% as the training set and 20% as the validation set.

### Implementation Details

For all benchmarks, we performed a leave-one-domain-out evaluation. We have integrated our advancements into three state-of-the-art algorithms for domain generalization, namely DG_via_ER (Zhao et al. 2020), EISNet (Wang et al. 2020), and FACT (Xu et al. 2021). These were chosen to allow for a comprehensive comparative evaluation. To maintain authenticity and fairness in comparison, we adhered to the parameter configurations presented in the original publications and their corresponding source code.

As an illustration, we incorporated a maximum loss into DG_via_ER's classification loss calculation, which originally used cross-entropy. Following this modification, we utilized SAM in conjunction with the original optimizer to update parameters based on the computed gradient of the classification loss.

For all experiments, we employed the SGD optimizer with a momentum and decay rate set at 0.9 and 0.0005, respectively. The learning rate was kept at 0.001. For our proposed enhancements, which we denote as SAM, the base optimizer was set to SGD, with rho at 0.1, learning rate at 0.01, adaptive set to False, weight decay at 0.0005, momentum at 0.9, and nesterov enabled. Concurrently, the weight of Maximum Square Loss was represented by $\gamma$ and set as 1 in the comparison experiment. For a more detailed expla-

| Algorithm | PACS | Officehome | DigitDG | DomainNet | AVG |
|---|---|---|---|---|---|
| ResNet18 | | | | | |
| DG_via_ER | 81.32 | 63.17 | 80.14 | 39.71 | 61.15 |
| DG_via_ER+Ours | **83.58** | **64.15** | **83.12** | **42.11** | **63.53** |
| EISNet | 81.77 | 63.47 | 80.38 | 38.44 | 60.75 |
| EISNet+Ours | **83.24** | **66.26** | **83.24** | **39.97** | **63.70** |
| FACT | 84.29 | 61.89 | 80.38 | 43.63 | 62.80 |
| FACT+Ours | **85.36** | **64.53** | **83.24** | **45.22** | **64.94** |
| ResNet50 | | | | | |
| DG_via_ER | 85.26 | 66.03 | 80.14 | 42.32 | 65.55 |
| DG_via_ER+Ours | **87.36** | **68.17** | **83.12** | **44.67** | **68.29** |
| EISNet | 85.64 | 66.23 | 80.38 | 44.80 | 65.98 |
| EISNet+Ours | **87.27** | **68.14** | **83.24** | **46.52** | **68.95** |
| FACT | 87.94 | 66.92 | 81.34 | 49.53 | 67.96 |
| FACT+Ours | **90.08** | **69.29** | **83.11** | **50.97** | **69.88** |

Table 1: Results(%) of our method combine other baselines.

| Algorithm | Art Painting | Cartoon | Photo | Sketch | AVG |
|---|---|---|---|---|---|
| ResNet18 | | | | | |
| DG_via_ER | 81.21± 0.47 | 76.20±0.45 | 96.15±0.27 | 71.75±1.09 | 81.32 |
| DG_via_ER +Ours | **83.25±0.19** | **79.30±0.12** | **97.08±0.10** | **74.70±0.18** | **83.58** |
| EISNet | 81.77±1.26 | 76.40±0.32 | 94.71±0.10 | 74.36±0.86 | 81.81 |
| EISNet+Ours | **83.42±0.56** | **77.57±0.33** | **95.89±0.16** | **77.47±0.43** | **83.59** |
| FACT | 84.59±0.59 | 78.17±0.26 | 95.15±0.10 | 79.23±0.19 | 84.29 |
| FACT+Ours | **85.30±0.23** | **79.49±0.42** | **96.47±0.15** | **80.17±0.25** | **85.36** |
| ResNet50 | | | | | |
| DG_via_ER | 87.39±1.09 | 79.31±1.40 | 98.04±0.17 | 76.30±0.65 | 85.26 |
| DG_via_ER +Ours | **89.25±0.53** | **82.07±0.86** | **98.33±0.11** | **79.79±0.44** | **87.36** |
| EISNet | 86.01±0.61 | 81.37±0.74 | 97.29±0.21 | 77.89±0.46 | 85.64 |
| EISNet+Ours | **87.94±0.29** | **82.42±0.41** | **98.11±0.17** | **80.61±0.74** | **87.27** |
| FACT | 89.53±0.72 | 81.49±0.22 | 96.69±0.08 | 84.03±0.54 | 87.94 |
| FACT+Ours | **90.55±0.27** | **84.32±0.55** | **97.93±0.17** | **87.83±0.27** | **90.08** |

Table 2: Leave-one-domain-out results(%) on PACS.

| Algorithm | Art | Clipart | Product | Realworld | AVG |
|---|---|---|---|---|---|
| DG_via_ER | 61.19±0.19 | 52.79±0.84 | 74.53±0.19 | 75.59±0.33 | 66.03 |
| DG_via_ER +Ours | **62.32±0.42** | **55.17±0.19** | **76.21±0.25** | **78.97±0.27** | **68.17** |
| EISNet | 62.59±0.71 | 53.19±0.14 | 73.97±0.32 | 75.17±0.19 | 66.23 |
| EISNet+Ours | **65.19±0.72** | **55.41±0.27** | **74.93±0.21** | **77.01±0.34** | **68.14** |
| FACT | 61.03±0.62 | 55.73±0.34 | 74.52±0.76 | 76.41±0.72 | 66.92 |
| FACT+Ours | **64.42±0.52** | **57.71±0.71** | **76.09±0.17** | **78.92±0.57** | **69.29** |

Table 3: Leave-one-domain-out results(%) on Officehome.

| Algorithm | MNIST | MNIST-M | SVYN | SYN | AVG |
|---|---|---|---|---|---|
| DG_via_ER | 96.93±0.23 | 63.79±0.57 | 71.04±0.76 | 88.79±0.27 | 80.14 |
| DG_via_ER +Ours | **98.13±0.11** | **69.43±0.33** | **74.09±0.31** | **90.81±0.17** | **83.12** |
| EISNet | 96.42±0.31 | 64.15±0.29 | 71.54±0.35 | 89.42±0.19 | 80.38 |
| EISNet+Ours | **97.79±0.26** | **70.23±0.32** | **73.54±0.29** | **91.41±0.12** | **83.24** |
| FACT | 97.63±0.13 | 65.23±0.47 | 72.22±1.06 | 90.27±0.13 | 81.34 |
| FACT+Ours | **98.71±0.34** | **68.18±0.26** | **74.01±0.29** | **91.53±0.20** | **83.11** |

Table 4: Leave-one-domain-out results(%) on Digits-DG.

nation of parameter selection, please refer to the 'Parameter Analysis' section. As the backbone of our model, we utilized ResNet18 and ResNet50 (He et al. 2016), the most commonly used networks in the field.

## Comparing Experimental Result

To verify the effectiveness of our proposed method, we tested it on three other baselines including DG_via_ER (Zhao et al. 2020), EISNet (Wang et al. 2020), and FACT (Xu et al. 2021) in Table. 1. These experiments were carried out on four distinct datasets, employing both ResNet18 and ResNet50 as the backbone networks. Due to space constraints, the average accuracy of each dataset is shown here.

It can be seen that when ResNet18 is used as the backbone, the average precision of DG_via_ER+Ours, EISNet+Ours, and FACT+Ours increases by 2.38%, 2.95% and 2.14% respectively. When ResNet50 was used as the backbone, the average accuracy of DG_via_ER+Ours, EISNet+Ours, and FACT+Ours increased by 2.74%,2.97%, and 1.92% respectively. All the above comparisons not only demonstrate the effectiveness of our proposed method but also shows that our method is a plug-and-play method.

| Settings | Value | AVG |
|---|---|---|
| SAM-Base Optimizer | SGD | **83.58** |
| | Adam | 64.59 |
| SAM-rho | 0.01 | **83.58** |
| | 0.1 | 75.53 |
| SAM-adaptive | False | **83.58** |
| | True | 80.14 |
| SAM-nesterov | False | 81.35 |
| | True | **83.58** |
| loss-$\gamma$ | 0.1 | 81.01 |
| | 0.5 | 82.05 |
| | 1 | **83.58** |

Table 5: Parameter analysis was performed on PACS with ResNet18.

## Experimental Analysis

**Ablation Experiment** In the ablation experiment, DG_via_ER is selected as the baseline, ResNet18 is used as the backbone and leave-one-domain-out evaluation is performed on PACS. The experimental results are shown in Table 6. The improvement of precision considering Smoothing Loss Landscape and Maximum Square Loss are compared respectively. In the ablation experiment, DG_via_ER did not make any improvement. To consider a Smoothing Loss Landscape, we add SAM to DG_via_ER and name it DG_via_ER+SAM, and the Maximum Square Loss is considered in DG_via_ER+loss. For each leave-one-domain-out evaluation, we conducted 5 experiments and obtained the average results.

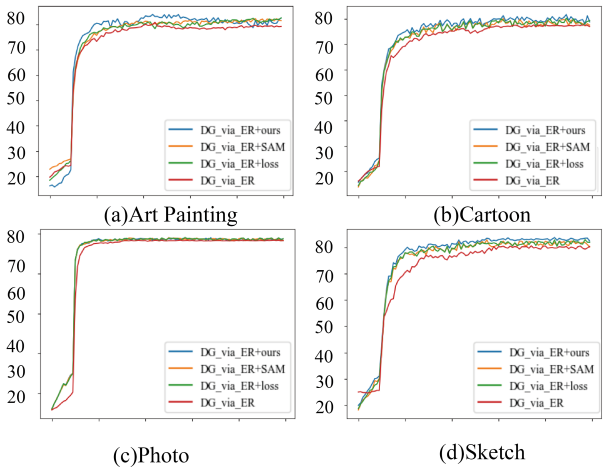



Figure 2: Accuracy curve of ablation experiment.

It can be seen that compared with the baseline, the average accuracy of DG_via_ER+SAM and DG_via_ER+loss increased by 0.99% and 1.38%, respectively. In addition, DG_via_ER+ours takes both of these improvements into account, thus obtaining a 2.26% accuracy improvement.

Meanwhile, the convergence curve of the model during the leave-one-domain-out evaluation in the ablation experiment can be seen in Figure 2. It can be clearly seen from Figure 2(b), and Figure 2(d) that after SAM and loss are used, the convergence speed of the model is also improved, and it will converge at around 30 epochs. However, the baseline gradually converges at around 60 epochs.

In conclusion, our ablation study highlights the effectiveness of augmenting the DG_via_ER algorithm with addi-

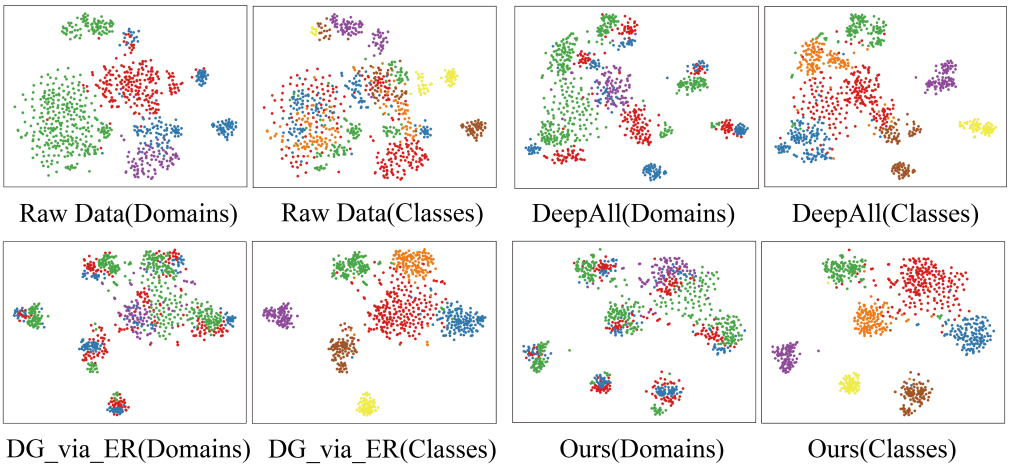



Figure 3: Feature visualization. Left: different colors represent different domains; Right: different colors indicate different classes.

| Algorithm | Art Painting | Cartoon | Photo | Sketch | AVG |
|---|---|---|---|---|---|
| DG_via_ER | 81.21± 0.47 | 76.20±0.45 | 96.15±0.27 | 71.75±1.09 | 81.32 |
| DG_via_ER+SAM | 81.79±0.40 | 77.52±0.29 | 96.92±0.08 | 72.99±0.35 | 82.31 |
| DG_via_ER+loss | 82.40±0.03 | 78.40±0.29 | 96.47±0.07 | 73.53±0.64 | 82.70 |
| DG_via_ER+Ours | **83.25±0.19** | **79.30±0.12** | **97.08±0.10** | **74.70±0.18** | **83.58** |

Table 6: Leave-one-domain-out results(%) on PACS. Ablation studies on the changes of Ours with ResNet18.

tional components such as SAM, loss, and our proposed method (Ours). The best overall performance is achieved by the DG_via_ER + Ours algorithm, with an average accuracy of 83.58. Future work could explore other possible enhancements to further improve the performance of the DG algorithm.

We use ResNet18 as the backbone to conduct the leave-one-domain-out evaluation in PACS to complete parameter analysis experiments of different weight factors to examine their impacts. We report the average accuracy of 5 trials in Table 5. In fact, for SAM, as long as the parameters are effective in helping the optimizer converge, our improvements should have some effect. For the weight factor $\gamma$ of Maximum Square Loss, when the value of $\gamma$ exceeds 0.5, the effect of domain generalization can be improved.

**Feature Visualization** To better understand the distribution of the learned features, we exploit t-SNE (Van der Maaten and Hinton 2008) to analyze the feature space learned by DeepAll, DG_via_ER, and DG_via_ER +Ours, respectively. DeepAll is a simple classification using ResNet18. We conduct this study on PACS; specifically, we take the Art Painting as the target, and others as the source. As shown in Figure 3, in the original feature space, the differences between domains outweigh the differences between categories. DeepAll has been able to distinguish different categories in the original feature space by simple classification, but the edges of the clusters are not obvious. Both Ours and DG_via_ER are capable of minimizing the distance between the distributions of the domains. Furthermore, the comparison between Ours (Classes, Domains) and DG_via_ER (Classes, Domains) can show that our method can distinguish data better.

## Conclusion

In this paper, we investigate the effect of smoothness-enhancing formulations on domain adversarial training, which combines task loss and adversarial loss objectives. Our approach is based on the idea that converging to a smooth minimum concerning task loss can stabilize the task loss and result in better performance on unseen domains. Moreover, we acknowledge that the distribution of objects in the real world often follows a power law, leading to a gap between machine learning models and our expectations of their performance on datasets with long-tailed class distributions. To handle this challenge, we approach the domain generalization problem from the angle of the long-tail distribution and suggest using the maximum square loss to balance different classes. We demonstrate the effectiveness of our method by comparing it with state-of-the-art methods on various domain generalization datasets.

## References

Balaji, Y.; Sankaranarayanan, S.; and Chellappa, R. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2229–2238.

Chen, M.; Xue, H.; and Cai, D. 2019. Domain Adaptation for Semantic Segmentation With Maximum Squares Loss. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2090–2099. IEEE.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Ghifary, M.; Balduzzi, D.; Kleijn, W. B.; and Zhang, M. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7): 1414–1430.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, 529–536.

Gulrajani, I.; and Lopez-Paz, D. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018a. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018b. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018c. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, 624–639.

Liu, Q.; Chen, C.; Dou, Q.; and Heng, P.-A. 2022. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1756–1764.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*, 10–18. PMLR.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Rangwani, H.; Aithal, S. K.; Mishra, M.; Jain, A.; and Radhakrishnan, V. B. 2022a. A Closer Look at Smoothness in Domain Adversarial Training. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 18378–18399. PMLR.

Rangwani, H.; Aithal, S. K.; Mishra, M.; and R., V. B. 2022b. Escaping Saddle Points for Effective Generalization on Class-Imbalanced Data. In *NeurIPS*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5018–5027.

Wan, C.; Shen, X.; Zhang, Y.; Yin, Z.; Tian, X.; Gao, F.; Huang, J.; and Hua, X.-S. 2022. Meta convolutional neural networks for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4682–4691.

Wang, M.; Li, P.; Shen, L.; Wang, Y.; Wang, S.; Wang, W.; Zhang, X.; Chen, J.; and Luo, Z. 2022a. Informative pairs mining based adaptive metric learning for adversarial domain adaptation. *Neural Networks*, 151: 238–249.

Wang, M.; Wang, S.; Wang, W.; Shen, L.; Zhang, X.; Lan, L.; and Luo, Z. 2023. Reducing bi-level feature redundancy for unsupervised domain adaptation. *Pattern Recognition*, 137: 109319.

Wang, M.; Wang, W.; Li, B.; Zhang, X.; Lan, L.; Tan, H.; Liang, T.; Yu, W.; and Luo, Z. 2021a. Interbn: Channel fusion for adversarial unsupervised domain adaptation. In *Proceedings of the 29th ACM international conference on multimedia*, 3691–3700.

Wang, M.; Yuan, J.; Qian, Q.; Wang, Z.; and Li, H. 2022b. Semantic data augmentation based distance metric learning for domain generalization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3214–3223.

Wang, S.; Yu, L.; Li, C.; Fu, C.-W.; and Heng, P.-A. 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, 159–176. Springer.

Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021b. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 834–843.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Yang, F.-E.; Cheng, Y.-C.; Shiau, Z.-Y.; and Wang, Y.-C. F. 2021. Adversarial teacher-student representation learning for domain generalization. *Advances in Neural Information Processing Systems*, 34: 19448–19460.

Yang, Y.; Wang, H.; and Katabi, D. 2022. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, 57–75. Springer.

Yao, X.; Bai, Y.; Zhang, X.; Zhang, Y.; Sun, Q.; Chen, R.; Li, R.; and Yu, B. 2022. PCL: Proxy-based Contrastive Learning for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7097–7107.

Zhang, G.; Zhao, H.; Yu, Y.; and Poupart, P. 2021. Quantifying and improving transferability in domain generalization. *Advances in Neural Information Processing Systems*, 34: 10957–10970.

Zhang, H.; Zhang, Y.-F.; Liu, W.; Weller, A.; Schölkopf, B.; and Xing, E. P. 2022a. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8024–8034.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022b. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8035–8045.

Zhao, S.; Gong, M.; Liu, T.; Fu, H.; and Tao, D. 2020. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems*, 33: 16096–16107.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2021. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, 561–578. Springer.