remote sensing

MDPI

Article

Implicit Sharpness-Aware Minimization for Domain Generalization

Mingrong Dong [1,2], Yixuan Yang [1,2], Kai Zeng [1,2], Qingwang Wang [1,2] and Tao Shen [1,2,*]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Wujiaying Street, Kunming 650500, China; dongmingrong@kust.edu.cn (M.D.); yangyixuan@stu.kust.edu.cn (Y.Y.); zengkai@kust.edu.cn (K.Z.); wangqingwang@kust.edu.cn (Q.W.)
[2] Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Wujiaying Street, Kunming 650500, China
* Correspondence: shentao@kust.edu.cn

Abstract: Domain generalization (DG) aims to learn knowledge from multiple related domains to achieve a robust generalization performance in unseen target domains, which is an effective approach to mitigate domain shift in remote sensing image classification. Although the sharpness-aware minimization (SAM) method enhances DG capability and improves remote sensing image classification performance by promoting the convergence of the loss minimum to a flatter loss surface, the perturbation loss (maximum loss within the neighborhood of a local minimum) of SAM fails to accurately measure the true sharpness of the loss landscape. Furthermore, its variants often overlook gradient conflicts, thereby limiting further improvement in DG performance. In this paper, we introduce implicit sharpness-aware minimization (ISAM), a novel method that addresses the deficiencies of SAM and mitigates gradient conflicts. Specifically, we demonstrate that the discrepancy in training loss during gradient ascent or descent serves as an equivalent measure of the dominant eigenvalue of the Hessian matrix. This discrepancy provides a reliable measure for sharpness. ISAM effectively reduces sharpness and mitigates potential conflicts between gradients by implicitly minimizing the discrepancy between training losses while ensuring a sufficiently low minimum through minimizing perturbation loss. Extensive experiments and analyses demonstrate that ISAM significantly enhances the model's generalization ability on remote sensing and DG datasets, outperforming existing state-of-the-art methods.

Keywords: domain generalization; sharpness-aware minimization; neural networks; loss landscape; remote sensing

Citation: Dong, M.; Yang, Y.; Zeng, K.; Wang, Q.; Shen, T. Implicit Sharpness-Aware Minimization for Domain Generalization. *Remote Sens.* **2024**, *16*, 2877. https://doi.org/10.3390/rs16162877

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 14 May 2024
Revised: 15 July 2024
Accepted: 19 July 2024
Published: 6 August 2024

1. Introduction

With the advancement of deep neural networks, deep learning techniques have achieved remarkable success in the field of the computer vision [1–4]. However, this success usually relies on the assumption that the training and test data have the same distribution. This enables an exceptional performance on training data (source domain), but the performance often degrades significantly when evaluated on test data (target domain), with distributions differing from the training data [5]. In remote sensing image acquisition, variations in geographic location, climate conditions, and sensors may exist. These factors give rise to domain shift phenomena for remote sensing images obtained at different times or locations, leading to changes in their distribution. Under such conditions, it is frequently necessary to retrain the model on data with different distributions to maintain its performance [6].

In order to ensure the model exhibits a robust performance across diverse data distributions, domain adaptation (DA) has been introduced as a solution for addressing out-of-distribution generalization [7,8]. DA necessitates access to well-labeled source

domain data and unlabeled target domain data during training, leveraging these unlabeled target domain data to reduce the gap between the source and target domains [9,10]. However, acquiring target domain data can be both costly and challenging in real-word scenarios [11]. For instance, remote sensing data collection not only demands expensive specialized equipment, but also requires skilled personnel for operation. Moreover, due to the wide range of environmental factors involved, it becomes arduous to encompass all potential situations. These challenges hinder the effective generalization of the model through DA in the absence of target domain data [12].

Domain generalization (DG) effectively addresses the limitations of relying on target domain data in DA [13]. DG does not necessitate any target domain data during training, providing a robust solution for scenarios where target domain data are either inaccessible or completely absent. DG enables the model to generalize to unseen domains by training the model across multiple accessible source domains [14]. Existing DG techniques comprise data manipulation [15,16], domain invariant representation learning [17,18], meta-learning [19–21], and gradient operations [22,23]. For example, Zhou et al. [24] probabilistically mix feature statistics to achieve data manipulation, and Hu et al. [5] extract domain invariant representations through an inter-domain alignment and inter-domain expansion. However, a study known as Domainbed [25] has revealed that most existing DG methods have an inferior performance compared to the empirical risk minimization (ERM) method under identical evaluation conditions. These failures have prompted the exploration of new methods.

A recent study, named SAM [26], proposed a method to ensure the loss minimum is located on a flat loss surface, which significantly enhances the model's DG performance. SAM introduces adversarial perturbations, ε, to the model parameters, θ, defining the loss at the point $\theta + \varepsilon$ as the maximum loss within the neighborhood of θ. By optimizing the perturbation loss, $\mathcal{L}(\theta + \varepsilon)$, instead of the original loss, $\mathcal{L}(\theta)$, this method aims to find a flat loss surface. However, it is important to note that both sharp and flat loss surfaces can achieve lower $\mathcal{L}(\theta + \varepsilon)$. The optimization objective employed by SAM does not accurately reflect the sharpness of the loss landscape, leading to minima that do not always converge to flat loss surfaces. Furthermore, most existing methods [27,28] primarily focus on optimizing the efficiency of SAM, neglecting this issue.

To address the issue present in SAM, Zhuang et al. [29] employed $h(\theta) = \mathcal{L}(\theta + \varepsilon) - \mathcal{L}(\theta)$ as an equivalent measure of sharpness and improved SAM by simultaneously optimizing $\mathcal{L}(\theta + \varepsilon)$ and $h(\theta)$. SAGM [30] further introduces an approach to simultaneously optimize $\mathcal{L}(\theta)$, $\mathcal{L}(\theta + \varepsilon)$, and $h(\theta)$ in order to find a flatter loss surface. However, these methods, which explicitly optimize combinations of $\mathcal{L}(\theta)$, $\mathcal{L}(\theta + \varepsilon)$, and $h(\theta)$, result in the parameter update direction becoming a linear combination of the gradients $\nabla\mathcal{L}(\theta)$ of the original loss and $\nabla\mathcal{L}(\theta + \varepsilon)$ of the perturbation loss. It is important to note that conflicts may arise between different gradients, and the gradients obtained at different parameter points may exacerbate conflicts. Consequently, the linear combination of $\nabla\mathcal{L}(\theta)$ and $\nabla\mathcal{L}(\theta + \varepsilon)$ acquired at different locations may leads to a deviation between the actual parameter update direction and the expected update direction, resulting in a suboptimal optimization outcome.

In this paper, we aim to address the issue of SAM and mitigate the adverse effects caused by gradient conflicts, thereby enhancing DG capability. We have demonstrated that the discrepancy in training loss values before and after executing gradient ascent or descent can serve as an equivalent measure of the dominant eigenvalue of the Hessian matrix. This discrepancy can be substituted for $h(\theta)$ as an equivalent measure of the sharpness of the loss landscape. As the logits vector output by the model leads to changes in the loss, the discrepancy between the logits vectors obtained in two trainings implies a discrepancy in the losses before and after training, thereby presenting information about the sharpness. Based on these findings, this paper introduces a new implicit measure of sharpness and a novel method named implicit sharpness-aware minimization (ISAM). ISAM avoids obtaining gradients at different locations, mitigating the adverse effects of gradient conflicts. Additionally, the implicit measure of sharpness prevents the issue in

SAM, where $\mathcal{L}(\theta + \varepsilon)$ fails to accurately reflect the sharpness, thus avoiding convergence to sharp regions.

In summary, our contributions are as follows:

Firstly, we examined the limitations of SAM and its variants in finding flat loss surfaces. Additionally, this paper explores how the discrepancy in training loss before and after executing a gradient ascent or descent can serve as an equivalent measure of the dominant eigenvalue of the Hessian matrix.

Secondly, we introduced a novel implicit measure of sharpness and an ISAM method, which mitigates the adverse effects caused by gradient conflicts between $\nabla\mathcal{L}(\theta)$ and $\nabla\mathcal{L}(\theta + \varepsilon)$ while improving on the inadequacy of SAM. This improvement significantly enhances the model's DG capabilities in unseen domains.

Finally, we conducted a series of experiments on several remote sensing datasets and DG datasets. The experimental results demonstrate that ISAM can effectively improve the generalization performance of the remote sensing image classification model, and the DG performance significantly outperforms current state-of-the-art DG methods.

2. Related Word

2.1. Domain Adaptation

DA enhances the generalization performance of the model across domains by narrowing the domain gap between the source and target domains [31]. Unsupervised DA, an effective approach within DA, improves the model's generalization capabilities during training by utilizing labeled source domain data and unlabeled target domain data [32,33]. Another pioneering approach, source-free DA, allows model tuning using solely unlabeled target domain data [34–36]. These techniques enhance the generalization performance by enabling the model to adapt to diverse scenarios across various environments. However, the acquisition of target domain data is not always feasible in real-life situations, thus limiting the scope of DA applications [37]. In contrast to DA, the goal of DG is to train a model to generalize to unseen domains in the complete absence of target domain data, relying only on source domain data. This positions DG closer to real-world conditions and makes it a more effective solution for practical scenarios.

2.2. Domain Generalization

Existing DG methods can be broadly categorized into data manipulation, representation learning, meta-learning, and gradient manipulation. Data manipulation enables the training space of the model to broaden to cover more unknown domains by increasing the number and diversity of source domain samples. Blanchard et al. [16] adjusted the distribution of training data to better approximate the distribution of data in unseen domains. Zhou et al. [38] created new domains by mixing training sample styles. Representation learning seeks to identify stable and invariant features across different domains and transfer these learned features to other domains to enhance DG performance. Kim et al. [39] employed contrastive regularization to encourage the model to learn representations. Nam et al. [40] removed style coding from a category prediction task to mitigate the effects of style differences. Meta-learning develops a generalized model by learning from previous experiences or tasks. Li et al. [21] used a meta-learning approach to simulate virtual target domains during training. Li et al. [20] improved the quality of feature representations through meta-learning strategies and guided the optimization of the feature extractor based on the learning output of the feature critic. Gradient manipulation utilizes gradient information to compel the neural network to learn a generalized representation. Shi et al. [22] optimized the training process by aligning gradients across different domains. However, the failure of most existing DG methods in Domainbed has motivated further explorations for new methods.

2.3. Sharpness-Aware Minimization

Recent research analyzing 40 different complexity measures has revealed a high correlation between sharpness-based metrics and model generalization capabilities [41]. This discovery

promotes an in-depth study of sharpness in model loss landscape. Foret et al. [26] introduced an efficient and generalized training method named SAM, which encourages the convergence of the minimum to a flat loss surface during training. Zhang et al. [42] introduced the concept of first-order flatness and identified a flat surface by examining the magnitude of gradients. Zhuang et al. [29] proposed GSAM, aimed at addressing the issue where SAM does not consistently converge the minimum to a flat loss surface. Wang et al. [30] introduced SAGM, utilizing a gradient matching technique, and successfully achieved a leading performance in DG tasks. Unfortunately, methods optimized for sharpness tend to improve SAM by explicitly weighing the relationship between original loss and perturbation loss. This overlooks potential conflicts between the gradients of the original loss and the perturbation loss, resulting in parameter updates in directions that diverge from the expected trajectory. In this paper, we propose a novel ISAM method to facilitate the convergence of the minimum to a flat loss surface while mitigating potential conflicts between gradients.

3. Methodology

3.1. Preliminaries

In this paper, we define the model with weight parameters, θ, denoted as $f(\cdot;\theta)$. Assume there are K different source domains, $D_S = \left\{ D_S^i | i = 1, 2, \cdots, K \right\}$, where each source domain, D_S^i, contains N_i sample-label pairs, $\left\{ (x_j^i, y_j^i) \right\}_{j=1}^{N_i}$. The source domains exhibit different joint distributions: $P_{xy}^i \neq P_{xy}^j$ for $1 \leq i \neq j \leq K$. The goal of DG is to train the model solely with source domains, enabling it to demonstrate a good generalization performance on any unseen target domain, $D_T = \left\{ x_j^i \right\}_{j=1}^{N_t}$, where N_t represents the number of samples in the target domain. The training loss, using empirical risk minimization (ERM) across all source domains, D_S, is defined as follows:

$$\mathcal{L}(\theta; D) = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{N_i} \sum_{j=1}^{N_i} \updownarrow \left(f\left(x_j^i; \theta \right), y_j^i \right) \tag{1}$$

where $\updownarrow(\cdot;\cdot)$ represents the loss function employed to measure the gap between the model's predictions and the actual labels, and we assume that $\mathcal{L}(\theta; D)$ is twice differentiable.

The conventional ERM method often results in minimum converging to a sharp region of the loss landscape when optimizing weight parameters. As shown in Figure 1, although these sharp regions exhibit significantly low loss on source domain data, they are highly sensitive to slight parameter variations. This sensitivity can hinder the model's generalization ability on unseen domain data. In contrast, SAM encourages the minimum to be located on a flat surface of the loss landscape through a two-step iterative process.

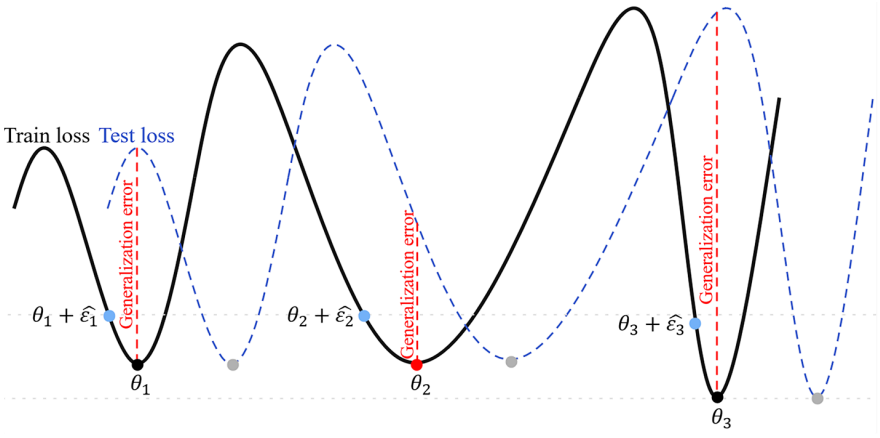


Figure 1. The case of different minima points in the loss landscape. Specifically, θ_3 situated within the sharp region exhibits a large generalization error, whereas θ_2 situated within the flatter region demonstrates a smaller generalization error.

The first iteration finds the point $\theta + \varepsilon$ where the loss is maximized within the Euclidean space centered at θ with a radius ρ, where $\rho \geq 0$, by using an adversarial approach:

$$\max_{\|\varepsilon\| \leq \rho} \mathcal{L}(\theta + \varepsilon; D) \tag{2}$$

where ε represents the perturbation to θ, and $\|\cdot\|$ represents the L_2 norm. In the second iteration, the gradient $\nabla \mathcal{L}(\theta + \varepsilon; D)$ is computed through backpropagation at $\theta + \varepsilon$, which is then utilized to update θ. Thus, the optimization objective of SAM can be formulated as follows:

$$\min_{\theta} \max_{\|\varepsilon\| \leq \rho} \mathcal{L}(\theta + \varepsilon; D) \tag{3}$$

The goal is to seek a flat loss surface by maximizing $\mathcal{L}(\theta + \varepsilon; D)$ followed by minimizing it. The approximate value of perturbation ε, denoted as $\hat{\varepsilon}$, can be approximated through a first-order Taylor expansion:

$$\begin{aligned} \hat{\varepsilon} = \underset{\|\varepsilon\| \leq \rho}{argmax} \mathcal{L}(\theta + \varepsilon; D) &\approx \underset{\|\varepsilon\| \leq \rho}{argmax} \mathcal{L}(\theta; D) + \varepsilon^T \nabla \mathcal{L}(\theta; D) \\ &= \rho \nabla \mathcal{L}(\theta; D) / \|\nabla \mathcal{L}(\theta; D)\| \end{aligned} \tag{4}$$

Ultimately, the perturbation loss of SAM can be formulated as:

$$\mathcal{L}_p(\theta; D) = \mathcal{L}(\theta + \hat{\varepsilon}; D) \tag{5}$$

3.2. SAM's Optimization and Gradient Conflict

Although SAM has demonstrated effectiveness in experiments, both sharp and flat minima regions can yield low perturbation loss values, indicating that perturbation loss is not always consistent with sharpness. As illustrated in Figure 1, despite the flatter loss surface at θ_2 compared to θ_1, $\mathcal{L}_p(\theta_2; D) = \mathcal{L}_p(\theta_1; D)$. This inconsistency may lead SAM to erroneously select θ_1 during the parameter selection process. Furthermore, although the perturbation loss at θ_3 is lower than that at θ_2, the gap between $\mathcal{L}(\theta_3; D)$ and $\mathcal{L}_p(\theta_3; D)$ is significantly greater than that between $\mathcal{L}(\theta_2; D)$ and $\mathcal{L}_p(\theta_2; D)$, prompting SAM to favor the sharper θ_3 over θ_2. It can be seen that SAM focuses too much on the perturbation loss in the training procedure, which results in a minimum that does not necessarily converge to a flat surface.

Fortunately, in the definition of SAM, the perturbation loss is defined as the maximum loss within a specified range, while original loss is considered a local minimum within that area. Therefore, the gap between the perturbation loss and the original loss can effectively serve as a measure of sharpness, as explained below:

$$h(\theta; D) = \mathcal{L}_p(\theta; D) - \mathcal{L}(\theta; D) \tag{6}$$

The function $h(\theta; D)$ more accurately describes the gap between θ and $\theta + \hat{\varepsilon}$, where a smaller gap indicates a flatter loss surface. By minimizing $h(\theta; D)$, it is possible to prevent the minimum from settling into a sharp loss region. Intuitively, optimizing both $\mathcal{L}_p(\theta; D)$ and $h(\theta; D)$, or $\mathcal{L}(\theta; D)$ and $h(\theta; D)$ simultaneously, becomes feasible. Minimizing $\mathcal{L}_p(\theta; D)$ ensures that the minimum point is sufficiently low, while minimizing $h(\theta; D)$ ensures that sharpness is also sufficiently small. As a result, it is possible to ensure convergence to a flat surface. The optimization objective is as follows:

$$\min_{\theta} \big(\mathcal{L}_p(\theta; D), h(\theta; D)\big) \tag{7}$$

The loss function can be formulated as $\alpha \mathcal{L}_p(\theta; D) + \beta h(\theta; D)$. Its gradient is as follows:

$$\begin{aligned} &\nabla \big(\alpha \mathcal{L}_p(\theta; D) + \beta h(\theta; D)\big) \\ &= \alpha \nabla \mathcal{L}_p(\theta; D) + \beta \big(\nabla \mathcal{L}_p(\theta; D) - \nabla \mathcal{L}(\theta; D)\big) \\ &= (\alpha + \beta) \nabla \mathcal{L}_p(\theta; D) - \beta \nabla \mathcal{L}(\theta; D) \end{aligned} \tag{8}$$

where α and β are positive scalars utilized to adjust the weight of the two objectives within the overall optimization.

Similarly, minimizing $\mathcal{L}(\theta; D)$ ensures the attainment of a low loss, and minimizing $h(\theta; D)$ guarantees that the minimum converges to a flat surface. The optimization objective is as follows:

$$\min_{\theta}(\mathcal{L}(\theta; D), h(\theta; D)) \tag{9}$$

Its gradient can be represented as follows:

$$\begin{aligned} &\nabla(\alpha\mathcal{L}(\theta; D) + \beta h(\theta; D)) \\ &= \alpha\nabla\mathcal{L}(\theta; D) + \beta\nabla\mathcal{L}_p(\theta; D) - \nabla\mathcal{L}(\theta; D) \\ &= (\alpha - \beta)\nabla\mathcal{L}(\theta; D) + \beta\nabla\mathcal{L}_p(\theta; D) \end{aligned} \tag{10}$$

Since $h(\theta; D)$ is closely related to $\mathcal{L}(\theta; D)$ and $\mathcal{L}_p(\theta; D)$, the method of explicitly combining $\mathcal{L}(\theta; D)$, $\mathcal{L}_p(\theta; D)$, and $h(\theta; D)$ eventually results in the gradient updating direction to be a linear combination of $\nabla\mathcal{L}(\theta; D)$ and $\nabla\mathcal{L}_p(\theta; D)$. Figure 2 illustrates the gradient update directions for $\mathcal{L}(\theta; D)$, $\mathcal{L}_p(\theta; D)$, and $h(\theta; D)$. As depicted in Figure 3, updating the gradient in a manner such as in Equations (8) and (10) would result in the direction of the update being significantly different from the directions of $-\nabla\mathcal{L}(\theta; D)$ and $-\nabla\mathcal{L}_p(\theta; D)$. This results in updating parameters that focus on decreasing the value of one optimization goal in one direction while increasing the value of the other optimization goal. Ultimately, the conflict between $\nabla\mathcal{L}(\theta; D)$ and $\nabla\mathcal{L}_p(\theta; D)$ manifests as follows: (1) the process of minimizing $\mathcal{L}(\theta; D)$ and $h(\theta; D)$ results in reducing $h(\theta; D)$ by increasing $\mathcal{L}(\theta; D)$; (2) the process of minimizing $\mathcal{L}_p(\theta; D)$ and $h(\theta; D)$ results in increasing $\mathcal{L}(\theta; D)$ to reduce the overall loss.

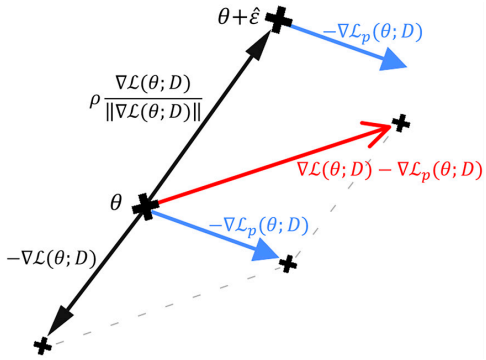


Figure 2. Illustration of the gradient update directions using different losses at parameter θ. The direction indicated by $-\nabla\mathcal{L}(\theta; D)$ is the gradient update direction of the original loss $\mathcal{L}(\theta; D)$, the direction indicated by $-\nabla\mathcal{L}_p(\theta; D)$ is the gradient update direction of the perturbation loss $\mathcal{L}_p(\theta; D)$, and the direction indicated by $\nabla\mathcal{L}(\theta; D) - \nabla\mathcal{L}_p(\theta; D)$ is the gradient update direction of $h(\theta; D)$.

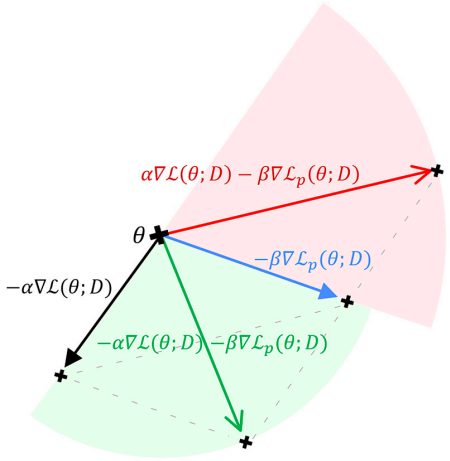


Figure 3. Potential directions for gradient updates: The pink region represents the possible directions of gradient updates when the optimization target is $\min_{\theta}(\mathcal{L}_p(\theta; D), h(\theta; D))$. The green region indicates the possible directions of gradient updates when the optimization target is $\min_{\theta}(\mathcal{L}(\theta; D), h(\theta; D))$.

3.3. Implicit Sharpness-Aware Minimization

As previously discussed, by explicitly optimizing the original loss $\mathcal{L}(\theta; D)$, the perturbation loss $\mathcal{L}_p(\theta; D)$, and $h(\theta; D)$, these actions ultimately translate into operations on $\nabla\mathcal{L}(\theta; D)$ and $\nabla\mathcal{L}_p(\theta; D)$. This optimization method had to face the adverse effects of conflicts between gradients. We must explore new methods to mitigate this drawback.

Assuming the loss function, $\mathcal{L}$, can be approximated by a second-order Taylor expansion near a local minimum, θ, we proceed with the following analysis. In the gradient descent process, if the same mini batch of samples, B, is used in iterations, we observe that the change in loss can be expressed as:

$$\mathcal{L}(\theta - \delta; D) = \mathcal{L}(\theta; D) - \delta^T \nabla\mathcal{L}(\theta; D) + \delta^T \nabla^2 \mathcal{L}(\theta; D)\delta/2 + O\left(\eta^3\right) \tag{11}$$

$$|R(\theta; D)| = |\mathcal{L}(\theta; D) - \mathcal{L}(\theta - \delta; D)| = \left|\delta^T \nabla\mathcal{L}(\theta; D) - \delta^T \nabla^2 \mathcal{L}(\theta; D)\delta/2 - O\left(\eta^3\right)\right| \tag{12}$$

where $\delta = \eta\nabla\mathcal{L}(\theta; D)$ represents the step size, η is the learning rate, and $|\cdot|$ denotes taking the absolute value. Around a local minimum, θ, $\nabla^2\mathcal{L}(\theta; D)$ denotes the Hessian matrix at θ. σ_{max} is defined as the dominant eigenvalue of the Hessian matrix. Consequently, we can derive:

$$|R(\theta; D)| = |\mathcal{L}(\theta; D) - \mathcal{L}(\theta - \delta; D)| \approx \left|\delta^T \nabla^2 \mathcal{L}(\theta; D)\delta/2\right| \tag{13}$$

$$\left|\sigma_{max}\left(\nabla^2 \mathcal{L}(\theta; D)\right)\right| \approx \left|2R(\theta; D)/\|\delta\|^2\right| \tag{14}$$

Kaur et al. [43] have demonstrated that σ_{max} can effectively measure the curvature at a minimum, allowing us to use $R(\theta; D)$ to replace $h(\theta; D)$ as an equivalent measure of sharpness.

Because changes in the logits vector can significantly influence the outcomes of the loss function calculations, we utilized the discrepancy between logits vectors to implicitly measure $R(\theta; D)$, thereby achieving an implicit measure of sharpness. For a mini batch $B = (x_1, x_2, \cdots, x_M)$ containing M samples, the logits vector $Z_B = (z_1, z_2, \cdots, z_M)$ generated by the model $f(\theta; D)$ and subsequently processed by the softmax function can be obtained as follows:

$$p(x_i) = \frac{exp(z_i)}{\sum_{j=1}^{M} exp(z_j)} \tag{15}$$

Predicted probabilities $p_{\theta-\delta}(x_i)$ and $p_\theta(x_i)$ are obtained from forward propagation at parameters $\theta - \delta$ and θ, respectively. By implicitly optimizing $R(\theta; D)$, we aim to optimize sharpness. Consequently, the following method can serve as an implicit measure of sharpness:

$$\frac{1}{M}\sum_{i=1}^{M} KL(p_{\theta-\delta}(x_i)\|p_\theta(x_i)) \tag{16}$$

Simultaneously obtaining both $p_{\theta-\delta}(x_i)$ and $p_\theta(x_i)$ necessitates an additional forward propagation step beyond SAM, leading to increased computation or memory consumption. We have discovered that, when performing a gradient ascent around a local minimum, θ, the discrepancy between losses can also serve as an equivalent measure of sharpness:

$$|\mathcal{L}(\theta; D) - \mathcal{L}(\theta - \delta; D)| = |\mathcal{L}(\theta + \delta; D) - \mathcal{L}(\theta; D)| \approx \left|\delta^T \nabla^2 \mathcal{L}(\theta; D)\delta/2\right| \tag{17}$$

Therefore, the discrepancy between $p_{\theta-\delta}(x_i)$ and $p_\theta(x_i)$, or $p_{\theta+\delta}(x_i)$ and $p_\theta(x_i)$, can serve as an implicit measure of sharpness. Before calculating $\nabla\mathcal{L}_p(\theta; D)$, SAM has already obtained the logits vector at the parameters $\theta + \hat{\varepsilon}$. Moreover, the direction of $\delta = \eta\nabla\mathcal{L}(\theta; D)$ aligns with that of $\hat{\varepsilon} = \rho\nabla\mathcal{L}(\theta; D)/\|\nabla\mathcal{L}(\theta; D)\|$, and both can be considered as increments for performing a gradient ascent at θ. To avoid additional computational overhead or memory consumption, we can equivalently substitute $p_{\theta+\hat{\varepsilon}}(x_i)$ obtained at the point $\theta + \hat{\varepsilon}$ for $p_{\theta+\delta}(x_i)$. Consequently, the implicit measure of sharpness term is defined as follows:

$$\frac{1}{M}\sum_{i=1}^{M} KL(p_{\theta+\hat{\varepsilon}}(x_i)\|p_\theta(x_i)) \tag{18}$$

In Figure 4, we demonstrate the impact of the implicit measure of sharpness term on sharpness. It can be observed that lower sharpness can be obtained with the use of the implicit measure of sharpness. Our final optimization objective is:

$$\mathcal{L}_{ISAM}(\theta; D) = \mathcal{L}_p(\theta; D) + \lambda \frac{1}{M} \sum_{i=1}^{M} KL(p_{\theta+\hat{\varepsilon}}(x_i) \| p_\theta(x_i)) \tag{19}$$

where λ is a hyperparameter used to balance the perturbation loss and the implicit measure of sharpness term.

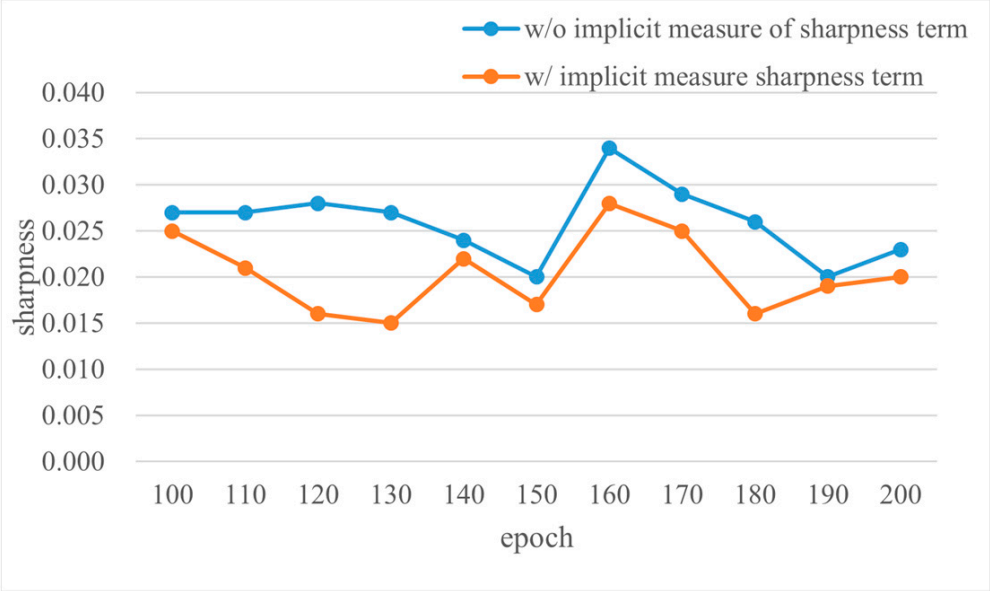


Figure 4. The impact of the implicit measure of sharpness term on sharpness. Sharpness is denoted by $h(\theta; D) = \mathcal{L}_p(\theta; D) - \mathcal{L}(\theta; D)$. The results were obtained using ResNet18 [44] on the SIRI-WHU dataset [45].

In order to address the shortcomings of SAM, previous work employed $h(\theta; D)$ as a measure of sharpness. A flat loss surface was sought by minimizing $h(\theta; D)$, and the convergence of the minimum to a lower loss surface was ensured by minimizing either $\mathcal{L}(\theta; D)$ or $\mathcal{L}_p(\theta; D)$. Ultimately, the parameters were updated using $\nabla\mathcal{L}(\theta; D)$ and $\nabla\mathcal{L}_p(\theta; D)$ obtained at θ and $\theta + \hat{\varepsilon}$. Since the two gradients are obtained at different parameter locations, these methods for finding a flat loss surface ignore and exacerbate the likelihood of conflicts between gradients. This leads to a suboptimal outcome by increasing $\mathcal{L}(\theta; D)$ or $\mathcal{L}_p(\theta; D)$ to minimize the overall loss.

In contrast to previous work, ISAM employs $\frac{1}{M} \sum_{i=1}^{M} KL(p_{\theta+\hat{\varepsilon}}(x_i) \| p_\theta(x_i))$ as an implicit measure of sharpness to address the issue that $\mathcal{L}_p(\theta; D)$ cannot accurately reflect sharpness in SAM. This implicit measure enhances the model's predictive consistency within the parameter space before and after perturbation, ensuring that the model's output remains relatively stable, even with slight changes in parameters. By minimizing this implicit sharpness measure, the minimum converges to a flat surface. Concurrently, ISAM ensures that the minimum converges to a surface with lower loss by minimizing $\mathcal{L}_p(\theta; D)$. Since both optimization objectives of ISAM compute the gradient at $\theta + \hat{\varepsilon}$, this greatly avoids exacerbating gradient conflicts. ISAM also mitigates the problem of previous work, where gradient conflicts between $\nabla\mathcal{L}(\theta; D)$ and $\nabla\mathcal{L}_p(\theta; D)$ resulted in increasing $\mathcal{L}(\theta; D)$ or $\mathcal{L}_p(\theta; D)$ to reduce the overall loss. Algorithm 1 describes the complete ISAM algorithm.

Algorithm 1: The Algorithm of ISAM

Input: Model f, source domains D_s, initial weight θ_0, learning rate η, training steps T, sample mini-batch $B \in D_s$, hyperparameter λ.
Output: Model trained with ISAM:

for $i \leftarrow 1$ **to** T **do**
  Compute the logits vector for the current parameters: $p_{\theta_t} = f(B; \theta_t)$;
  Calculate the original loss gradient: $\nabla \mathcal{L}(\theta_t; D)$;
  Calculate $\hat{\varepsilon}$ according to Equation (4);
  Compute the logits vector for perturbed parameters: $P_{\theta_t + \hat{\varepsilon}} = f(B; \theta_t + \hat{\varepsilon})$;
  $\mathcal{L}_{ISAM} = \mathcal{L}_p(\theta_t; D) + \lambda \frac{1}{M} \sum_{i=1}^{M} KL(p_{\theta_t + \hat{\varepsilon}}(x_i) \| p_{\theta_t}(x_i))$;
  Compute the gradient approximation of the ISAM objective: $g = \nabla \mathcal{L}_{ISAM}(\theta_t; D)|_{\theta_t + \hat{\varepsilon}}$
  Update weights: $\theta_{t+1} = \theta_t - \eta g$;
end for

4. Experiments

4.1. Experiment Setups and Implementation Details

4.1.1. Dataset

We conducted image classification tasks on four remote sensing datasets to evaluate the influence of our proposed method on the generalization performance of the models. The UC Merced Land-Use dataset [46] contains a total of 2100 images with dimensions of (3, 256, 256), spanning 21 categories. The SIRI-WHU dataset [45] contains a total of 2400 images with dimensions of (3, 200, 200), spanning 12 categories. The RSSCN7 dataset [47] contains a total of 2800 images with dimensions of (3, 400, 400), spanning 7 categories. The RSC11 dataset [48] contains a total of 1232 images with dimensions of (3, 400, 400), spanning 11 categories. Concurrently, due to the lack of a dedicated DG dataset within the remote sensing field, we conducted image classification tasks on 3 public datasets of DG to evaluate our proposed method and competing methods. The PACS dataset [49] contains a total of 9991 images with dimensions of (3, 224, 224), spanning 7 categories and 4 domains: Art, Cartoon, Photo, and Sketch. The VLCS dataset [50] includes a total of 10,729 images with dimensions of (3, 224, 224), covering 5 categories and 4 domains: Caltech101, LabelMe, SUN09, and VOC2007. The Office-Home dataset [51] consists of 15,588 images with dimensions of (3, 224, 224), across 65 categories and 4 domains: Art, Clipart, Product, and Real.

4.1.2. Evaluation Protocol

For remote sensing datasets, we used 80% of the dataset as the training set and 20% as the validation set. Four metrics were used to evaluate the performance: accuracy, precision, recall, and F1-score. For DG datasets, we followed the model selection strategy of training domain validation sets in DG, as per previous work [52]. The strategy splits the source domain into two parts, 80% and 20%, where 80% is used as the training set for training the model and 20% is used as the validation set for model selection. Ultimately, this strategy selects the model that performs best on the validation set of the source domain.

4.1.3. Implementation Details

For the remote sensing datasets, the pretrained ResNet18 and ResNet50 [44] models on the ImageNet [53] dataset were used as the backbones. All methods use the SGD optimizer as the base optimizer, with the learning rate set to 0.01, the batch size set to 32, and each training after 200 epochs. For the DG datasets, the pretrained ResNet18 model on the ImageNet dataset was used as the backbone. Hyperparameters were randomly searched from predefined distributions in the Domainbed benchmark: learning rates were chosen from $10^{Uniform(-5,-2)}$, batch sizes from $2^{Uniform(3.5,5)}$, weight decay from $10^{Uniform(-6,-2)}$, and dropout rate from $\{0, 0.1, 0.5\}$. Each training of the model went through 5000 iterations, with validation every 300 iterations.

4.1.4. Baseline

On the remote sensing datasets, we compared the proposed ISAM with empirical risk minimization (ERM) [54], sharpness-aware minimization (SAM) [26], and sharpness-aware gradient matching (SAGM) [30]. On the DG datasets, we compared the proposed method, ISAM, with conventional DG methods. The compared DG methods include: Mixup [55], MTL [16], Maximum Mean Discrepancy (MMD) [56], Class-conditional DANN (CDANN) [57], Adaptive Risk Minimization (ARM) [19], empirical risk minimization (ERM) [54], Group Distributionally Robust Optimization (GroupDRO) [58], Conditional Contrastive Adversarial Domain Bottleneck (CondCAD) [15], Deep CORrelation ALignment (CORAL) [59], Fish [22], Meta-Learning Domain Generalization (MLDG) [21], Fishr [23], Style-Agnostic Network (SagNet) [40], Self-Supervised Contrastive Regularization (SelfReg) [39], Variance Risk Extrapolation (VREx) [18], Spectral Decoupling (SD) [60], and Contrastive Adversarial Domain Bottleneck (CAD) [15].

4.2. Comparison Results

4.2.1. Comparison on Remote Sensing Datasets

We compared SAM, SAGM, and ISAM for their improvement of model generalization performance, and the results are shown in Table 1. In the SIRI-WHU and UC Merced Land-Use datasets, our approach achieved the best results on all four metrics. ISAM also demonstrated excellent performance on the RSC11 dataset. On ResNet18, ISAM's accuracy improved by 2.3%, 0.7%, 2%, and 1.2% compared to ERM across the four datasets. Similarly, on ResNet50, ISAM's accuracy improved by 1.7%, 0.9%, 2%, and 0.7% compared to ERM. ISAM exhibited a superior generalization performance improvement compared to SAM and SAGM. In Figure 5, we demonstrate the average accuracy, precision, recall, and F1-score of different methods on the four remote sensing datasets using ResNet18. Experimental results across these datasets indicate that ISAM can effectively enhance the generalization performance of the model in the remote sensing field.

Table 1. Comparison of ERM, SAM, SAGM, and ISAM on remote sensing datasets. The best and second-best results are indicated in bold and underlined, respectively.

		ResNet18				**ResNet50**			
Dataset	**Method**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
SIRI-WHU	ERM	95.4	95.4	95.3	95.3	96.6	96.7	96.5	96.6
	SAM	96.7	96.8	96.7	96.7	97.2	97.3	97.4	97.3
	SAGM	<u>96.9</u>	<u>96.9</u>	<u>96.9</u>	<u>96.9</u>	<u>97.9</u>	<u>98.0</u>	<u>97.8</u>	<u>97.9</u>
	ISAM (ours)	**97.7**	**97.6**	**97.7**	**97.7**	**98.3**	**98.2**	**98.2**	**98.2**
RSSCN7	ERM	96.4	96.4	96.4	96.4	96.8	96.9	96.9	96.8
	SAM	96.8	96.7	96.8	96.7	97.1	97.2	97.2	97.2
	SAGM	<u>97.0</u>	<u>96.8</u>	**97.3**	<u>97.0</u>	<u>97.3</u>	<u>97.4</u>	<u>97.3</u>	<u>97.3</u>
	ISAM (ours)	**97.1**	**97.1**	<u>97.2</u>	**97.2**	**97.7**	**97.7**	**97.8**	**97.7**
RSC11	ERM	96.0	96.4	96.2	96.1	97.2	97.5	97.4	97.4
	SAM	<u>97.2</u>	97.3	97.0	97.1	<u>98.8</u>	<u>98.8</u>	98.6	<u>98.7</u>
	SAGM	**98.0**	**98.4**	<u>97.6</u>	**98.0**	<u>98.8</u>	98.7	<u>98.8</u>	<u>98.7</u>
	ISAM (ours)	**98.0**	<u>97.8</u>	**97.8**	<u>97.8</u>	**99.2**	**99.0**	**99.3**	**99.1**
UC Merced Land-Use	ERM	97.4	97.6	97.6	97.5	98.6	98.5	98.6	98.5
	SAM	97.4	97.4	97.7	97.5	98.8	98.7	98.7	98.6
	SAGM	<u>97.9</u>	<u>98.0</u>	<u>97.8</u>	<u>97.8</u>	<u>99.0</u>	<u>99.0</u>	<u>99.0</u>	<u>99.0</u>
	ISAM (ours)	**98.6**	**98.4**	**98.4**	**98.4**	**99.3**	**99.2**	**99.2**	**99.2**

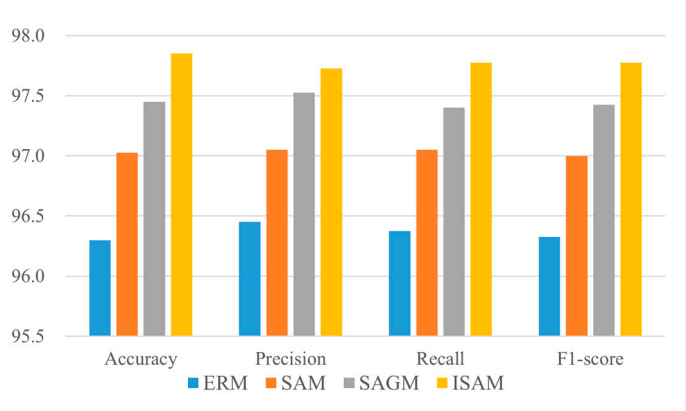


Figure 5. Average accuracy, precision, recall, and F1-score comparison of ISAM with ERM, SAM, and SAGM on four remote sensing datasets.

4.2.2. Comparison on DG Datasets

Table 2 reports the results on the PACS dataset. Our method achieved the best results in the Photo domain and in terms of average accuracy, also outperforming the ERM baseline across all four domains. Additionally, we conducted a visual comparison between ISAM and ERM using Grad-CAM [61] in Figure 6. The Grad-CAM heatmaps show that ISAM is able to better focus on important feature regions of the images while covering a broader area. As detailed in Table 3, our method excelled in the Caltech, VOC, and Sun domains of the VLCS dataset and achieved superior average accuracy. Its performance was only marginally lower than the ERM method by 0.2% in the LabelMe domain. Moreover, in Table 3, we note that the performance of the ERM method exceeds more than half of the compared methods. This phenomenon further demonstrates that methods that converge the minimum to a flat surface can significantly enhance DG performance compared to other approaches. The results for the Office-Home dataset are shown in Table 4. Our method still outperforms the other compared DG methods and achieved the best results in the Clipart and Real domains. Collectively, the results across all three datasets demonstrate the effectiveness of our proposed method in enhancing DG performance.

Table 2. Comparison with state-of-the-art domain generalization methods on PACS dataset. The best and second-best results are indicated in bold and underlined, respectively. The results marked with † are from [52].

Method	Art	Cartoon	Photo	Sketch	Average
Mixup	80.7	71.7	94.6	71.3	79.6
MTL †	78.7	73.4	94.1	74.4	80.2
MMD	77.9	76.7	94.2	71.9	80.2
CDANN †	80.4	73.7	93.1	74.2	80.4
ARM †	79.4	75.0	94.3	73.8	80.6
ERM	78.7	74.4	95.1	74.7	80.7
GroupDRO †	77.7	76.4	94.0	74.8	80.7
CondCAD †	79.7	74.2	94.6	74.8	80.8
CORAL	79.7	**77.4**	93.8	73.7	81.2
Fish	77.7	77.1	94.5	75.5	81.2
MLDG †	78.4	75.1	94.8	<u>76.7</u>	81.3
Fishr †	81.2	75.8	94.3	73.8	81.3
SagNet †	82.9	73.2	94.6	76.1	81.7
SelfReg	82.5	74.4	<u>95.4</u>	74.9	81.8
VREx	78.8	<u>75.4</u>	94.0	**79.2**	<u>81.9</u>
SD †	<u>83.2</u>	74.6	94.6	75.1	<u>81.9</u>
CAD †	**83.9**	74.2	94.6	75.0	<u>81.9</u>
ISAM (ours)	82.6	74.8	**95.8**	75.5	**82.2**

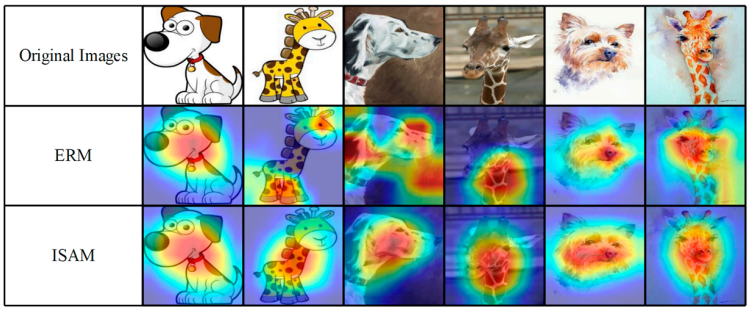


Figure 6. The Grad-CAM heatmap visualization. The results were obtained using ResNet18 on the PACS dataset.

Table 3. Comparison with state-of-the-art domain generalization methods on VLCS dataset. The best and second-best results are indicated in bold and underlined, respectively. The results marked with † are from [52].

Method	Caltech	LabelMe	Sun	VOC	Average
MMD	96.0	64.3	68.5	70.8	74.9
MTL †	94.4	**65.0**	69.6	71.7	75.2
MLDG †	95.8	63.3	68.5	73.1	75.2
CAD †	94.5	<u>63.5</u>	70.4	72.4	75.2
VREx	96.2	62.5	69.3	73.1	75.3
GroupDRO †	96.7	61.7	70.2	72.9	75.4
SagNet	94.9	61.9	69.6	75.2	75.4
SD †	96.5	62.2	69.7	73.6	75.5
CORAL	96.5	62.8	69.1	73.8	75.6
ERM	<u>97.7</u>	62.1	70.3	73.2	75.8
ARM †	96.9	61.9	<u>71.6</u>	73.3	75.9
CDANN †	95.4	62.6	69.9	<u>76.2</u>	76.0
CondCAD †	96.5	62.6	69.1	76.0	76.1
Fishr †	97.2	63.3	70.4	74.0	76.2
Mixup	95.6	62.7	71.3	75.4	76.3
SelfReg	95.8	63.4	71.1	75.3	76.4
Fish	97.4	63.4	71.5	75.2	<u>76.9</u>
ISAM (ours)	**98.7**	61.9	**71.8**	**77.9**	**77.6**

Table 4. Comparison with state-of-the-art domain generalization methods on Office-Home dataset. The best and second-best results are indicated in bold and underlined, respectively. The results marked with † are from [52].

Method	Art	Clipart	Product	Real	Average
VREx	49.2	46.2	68.0	68.0	57.9
MMD	49.2	46.7	69.4	70.3	58.9
CDANN †	51.4	46.9	68.4	70.4	59.3
ARM †	51.3	48.5	68.0	70.6	59.6
MTL †	51.6	47.7	69.1	71.0	59.9
ERM	52.1	47.1	70.0	70.5	59.9
CAD †	52.1	48.3	69.7	71.9	60.5
GroupDRO †	52.6	48.2	69.9	71.5	60.6
Fishr †	52.6	48.6	69.9	72.4	60.9
MLDG †	53.1	48.4	70.5	71.7	60.9
CondCAD †	53.3	48.4	69.8	72.6	61.0
Fish	55.6	49.1	71.4	71.7	62.0
SagNet	**56.5**	49.6	70.6	72.2	62.2
SelfReg	55.1	49.2	<u>72.2</u>	73.0	62.4
Mixup	<u>55.9</u>	49.8	71.6	72.4	62.4
SD †	55.0	51.3	**72.5**	72.7	62.9
CORAL	55.4	<u>51.5</u>	71.8	<u>73.2</u>	<u>63.0</u>
ISAM (ours)	55.8	**52.2**	<u>72.2</u>	**73.4**	**63.4**

4.3. Ablation Study and Parameter Analysis

4.3.1. Ablation Study on DG Datasets

To clearly demonstrate the capability of the proposed ISAM method in finding flat loss surfaces, we conducted ablation experiments with ERM, SAM, SAGM, and ISAM. The experimental results are detailed in Tables 5–7. Experiments across three datasets showed that the average accuracy of the SAM, SAGM, and ISAM methods all exceeded that of the ERM method. This indicates that methods enabling the minimum to converge to flat loss surfaces can effectively enhance DG performance. Compared to SAM, the improvements made by ISAM are significant. Apart from the Sun domain in the VLCS dataset, ISAM outperformed SAM across all domains and in average accuracy. Figure 7 demonstrates the feature clustering effects of SAM and ISAM, and it can be observed that ISAM has a better clustering effect.

Table 5. The ablation study of ISAM on the PACS dataset. The best and second-best results are indicated in bold and underlined, respectively. All results were obtained by a hyperparameter search of Domainbed.

Method	Art	Cartoon	Photo	Sketch	Average
ERM	78.7	74.4	95.1	74.7	80.7
SAM	81.0	74.2	95.1	74.8	81.3
SAGM	<u>81.6</u>	<u>74.3</u>	<u>95.3</u>	<u>75.1</u>	<u>81.6</u>
ISAM	**82.6**	**74.8**	**95.8**	**75.5**	**82.2**

Table 6. The ablation study of ISAM on the VLCS dataset. The best and second-best results are indicated in bold and underlined, respectively. All results were obtained by a hyperparameter search of Domainbed.

Method	Caltech	LabelMe	Sun	VOC	Average
ERM	97.7	<u>62.1</u>	70.3	73.2	75.8
SAM	<u>98.6</u>	60.7	**72.0**	76.8	77.0
SAGM	97.8	**62.6**	71.0	<u>77.0</u>	<u>77.1</u>
ISAM	**98.7**	61.9	<u>71.8</u>	**77.9**	**77.6**

Table 7. The ablation study of ISAM on the Office-Home dataset. The best and second-best results are indicated in bold and underlined, respectively. All results were obtained by a hyperparameter search of Domainbed.

Method	Art	Clipart	Product	Real	Average
ERM	52.1	47.1	70.0	70.5	59.9
SAM	54.1	49.5	**72.2**	**73.6**	62.4
SAGM	<u>54.3</u>	<u>49.8</u>	<u>72.1</u>	**73.6**	<u>62.5</u>
ISAM	**55.8**	**52.2**	**72.2**	<u>73.4</u>	**63.4**

Table 8 shows the optimization objectives of different methods. Compared to SAGM, which has the optimization objective of $\mathcal{L}(\theta; D) + \mathcal{L}_p(\theta - \lambda \nabla_\theta \mathcal{L}(\theta; D); D)$, ISAM leads by 0.6%, 0.5%, and 0.9% in accuracy on the PACS, VLCS, and Office-Home datasets, respectively. This fully demonstrates that the implicit measure of sharpness can effectively mitigate the adverse effects caused by conflicts between $\nabla \mathcal{L}(\theta; D)$ and $\nabla \mathcal{L}_p(\theta; D)$.

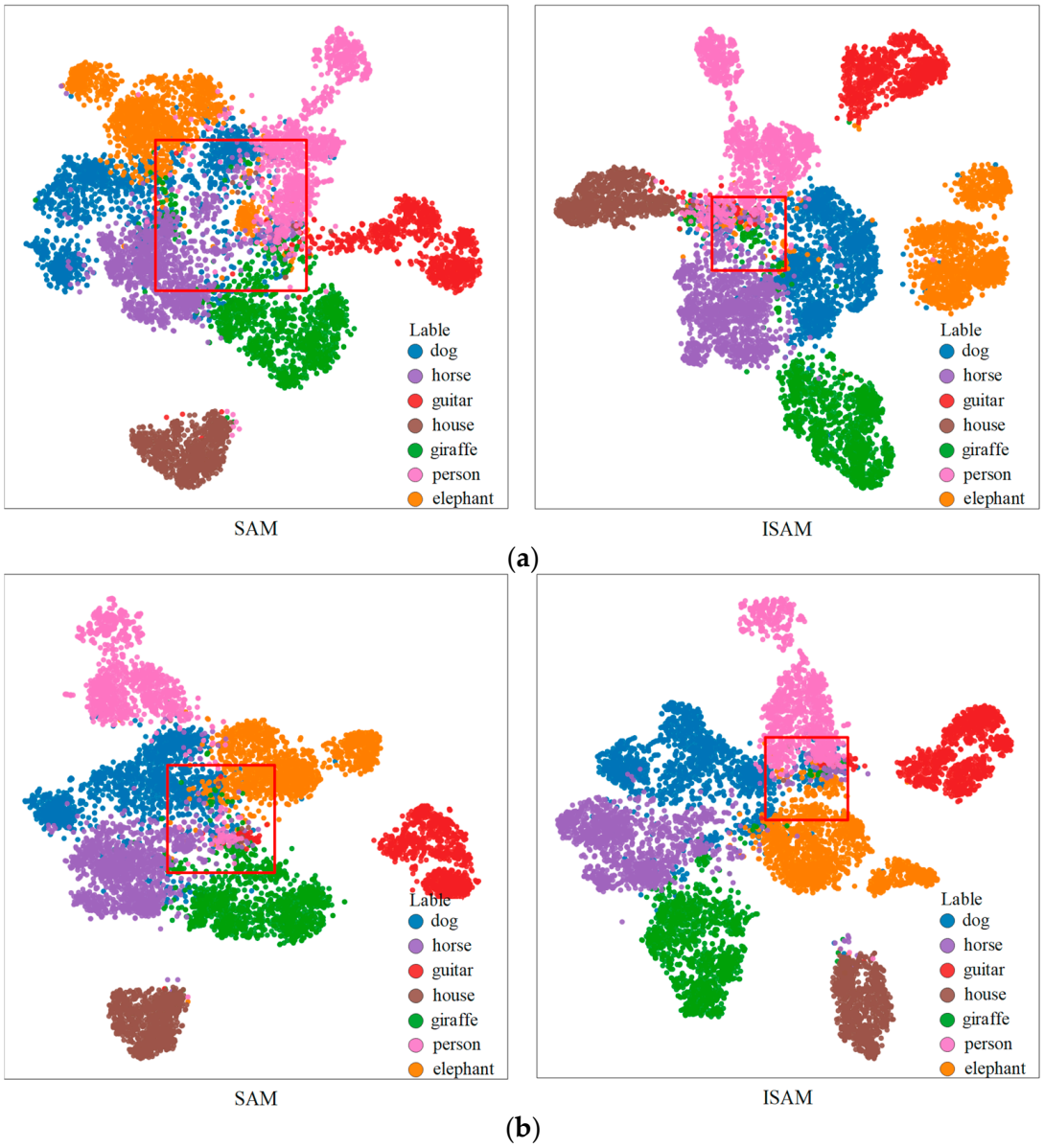


Figure 7. The t-SNE [62] visualization shows the clustering effects of SAM and ISAM. The results were obtained using ResNet18 on the PACS dataset. (**a**) Target domain: Art. (**b**) Target domain: Photo.

Table 8. Ablation study of ISAM. Optimization objectives of different methods and their DG accuracy (%).

Method	Optimization Objective	PACS	VLCS	Office-Home
ERM	$\mathcal{L}(\theta; D)$	80.7	75.8	59.9
SAM	$\mathcal{L}_p(\theta; D)$	81.3	77.0	62.4
SAGM	$\mathcal{L}(\theta; D) + \mathcal{L}_p(\theta - \lambda \nabla_\theta \mathcal{L}(\theta; D); D)$	81.6	77.1	62.5
ISAM	$\mathcal{L}_p(\theta; D) + \lambda \frac{1}{M} \sum_{i=1}^{M} KL(p_{\theta+\hat{\varepsilon}}(x_i) \Vert p_\theta(x_i))$	82.2	77.6	63.4

4.3.2. Ablation Study on CIFAR-10

We conducted ablation experiments on the CIFAR-10 [63] dataset using ResNet18 to compare ERM, SAM, SAGM, and ISAM. As depicted in Figure 8, ISAM demonstrated superior performance among these four methods. Although SAGM, an improved version of SAM, outperformed ERM, its accuracy was slightly lower than that of SAM. These results

further validate that ISAM effectively mitigates the adverse effects of conflicts between $\nabla\mathcal{L}(\theta;D)$ and $\nabla\mathcal{L}_p(\theta;D)$ while improving SAM.

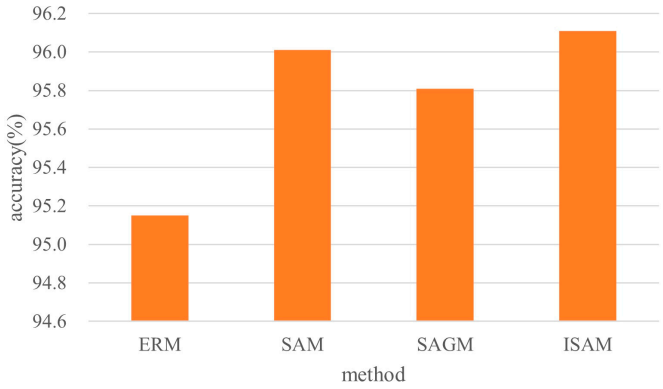
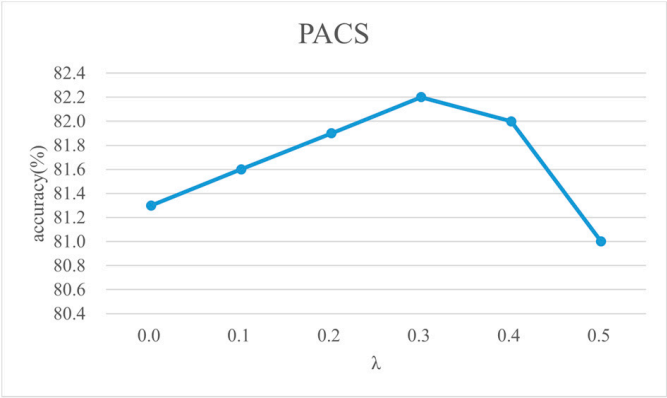


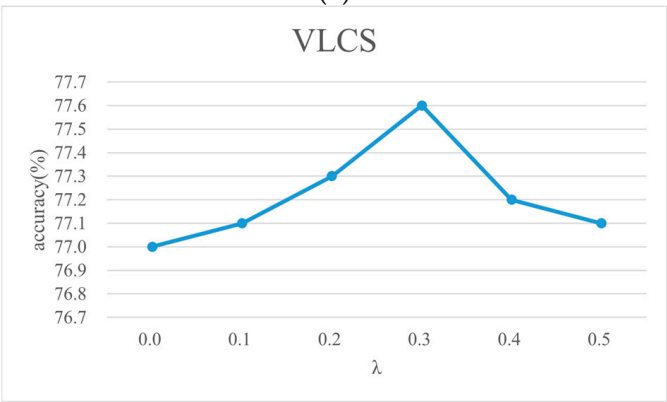
Figure 8. Classification accuracy of ERM, SAM, SAGM, and ISAM evaluated by ResNet18 on CIFAR10.

4.3.3. Parameter Analysis

We conducted a study on the hyperparameter, λ, for the implicit measure of sharpness term. λ is set to (0, 0.1, 0.2, 0.3, 0.4, 0.5). The experiments were carried out with all other parameters remaining constant, and the sequence of data used in each trial was fixed. Figure 9 illustrates the impact of different λ values on the generalization performance within the PACS, VLCS, and Office-Home datasets. It can be observed that the addition of the implicit measure of sharpness term effectively enhances DG performance.

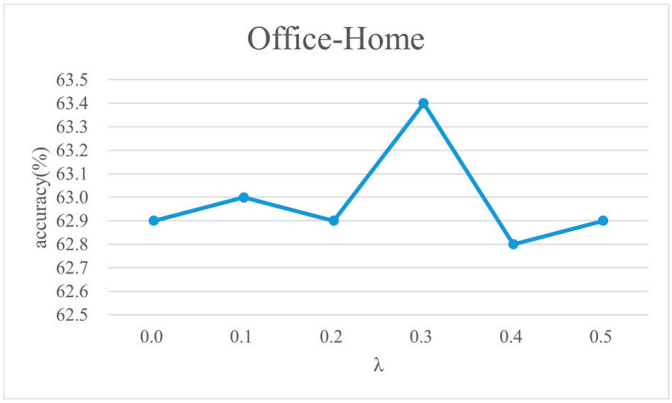


(**a**)



(**b**)

Figure 9. *Cont.*



(**c**)

Figure 9. The impact of the hyperparameter λ of the implicit measure of sharpness term on DG performance in ISAM. Results on different datasets with varying λ values using ResNet18: (**a**) PACS; (**b**) VLCS; (**c**) Office-Home.

5. Discussion

In this work, we introduce ISAM by proposing a novel measure of sharpness. ISAM addresses the limitations of SAM and mitigates the adverse effects caused by gradient conflicts. We validated the effectiveness of ISAM through extensive experiments on various remote sensing and DG datasets. Although ISAM has made progress in terms of DG performance, it has not reduced computational complexity during optimization. Consequently, the computational time and memory consumption increase when handling large-scale datasets. Our future research will focus on optimizing performance while incorporating more efficient computational techniques to reduce computational costs.

6. Conclusions

This paper provides a detailed analysis of the issues encountered by SAM and its variants in optimizing the sharpness of the loss landscape. In particular, we analyze the adverse effects resulting from gradient conflicts between the original loss and the perturbation loss. To mitigate these issues, we introduce an implicit measure of sharpness. Subsequently, we propose an algorithm named ISAM, which promotes the convergence of the loss minimum to a flat loss surface by minimizing the perturbation loss and the implicit measure of sharpness. ISAM effectively mitigates the adverse effects of conflicts between gradients while improving SAM. The extensive experiments on several remote sensing and DG datasets demonstrate that ISAM effectively enhances the model's DG performance.

Author Contributions: Conceptualization, M.D., Y.Y. and T.S.; methodology, M.D. and Y.Y.; software, Y.Y.; validation, Y.Y.; formal analysis, Y.Y.; resources, M.D., K.Z., Q.W. and T.S.; data curation, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, M.D.; visualization, Y.Y.; supervision, T.S.; project administration, Y.Y.; funding acquisition, M.D., K.Z. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the Yunnan Fundamental Research Projects under Grants 202301AV070003 and 202101BE070001-008, and in part by the Major Science and Technology Projects in Yunnan Province under Grants 202302AG050009 and 202202AD080013.

Data Availability Statement: Data for this article can be obtained by contacting the author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aggarwal, K.; Singh, S.K.; Chopra, M.; Kumar, S.; Colace, F. Deep learning in robotics for strengthening industry 4.0.: Opportunities, challenges and future directions. In *Robotics and AI for Cybersecurity and Critical Infrastructure in Smart Cities*; Springer: Cham, Switzerland, 2022; pp. 1–19. [CrossRef]
2. Jiang, W.; Yang, H.; Zhang, Y.; Kwok, J. An adaptive policy to employ sharpness-aware minimization. *arXiv* **2023**. [CrossRef]
3. Tsuneki, M. Deep learning models in medical image analysis. *J. Oral Biosci.* **2022**, *64*, 312–320. [CrossRef] [PubMed]
4. Yang, B.; Wang, C.; Ma, X.; Song, B.; Liu, Z.; Sun, F. Zero-Shot Sketch-Based Remote-Sensing Image Retrieval Based on Multi-Level and Attention-Guided Tokenization. *Remote Sens.* **2024**, *16*, 1653. [CrossRef]
5. Hu, J.; Qi, L.; Zhang, J.; Shi, Y. Domain generalization via Inter-domain Alignment and Intra-domain Expansion. *Pattern Recognit.* **2024**, *146*, 110029. [CrossRef]
6. Zhang, X.; Chen, Y.C. Adaptive Domain Generalization Via Online Disagreement Minimization. *IEEE Trans. Image Process.* **2023**, *32*, 4247–4258. [CrossRef] [PubMed]
7. Xu, Q.; Zhang, R.; Fan, Z.; Wang, Y.; Wu, Y.-Y.; Zhang, Y. Fourier-based augmentation with applications to domain generalization. *Pattern Recognit.* **2023**, *139*, 109474. [CrossRef]
8. Guan, H.; Liu, M. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1173–1185. [CrossRef] [PubMed]
9. Shao, Y.; Li, L.; Ren, W.; Gao, C.; Sang, N. Domain adaptation for image dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2808–2817.
10. Hsu, H.-K.; Yao, C.-H.; Tsai, Y.-H.; Hung, W.-C.; Tseng, H.-Y.; Singh, M.; Yang, M.-H. Progressive domain adaptation for object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 2–5 March 2020; pp. 749–757.
11. Niu, Z.; Yuan, J.; Ma, X.; Xu, Y.; Liu, J.; Chen, Y.W.; Tong, R.; Lin, L. Knowledge Distillation-based Domain-invariant Representation Learning for Domain Generalization. *IEEE Trans. Multimed.* **2023**, *26*, 245–255. [CrossRef]
12. Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; Loy, C.C. Domain Generalization: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4396–4415. [CrossRef] [PubMed]
13. Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; Yu, P.S. Generalizing to Unseen Domains: A Survey on Domain Generalization. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 8052–8072. [CrossRef]
14. Eastwood, C.; Robey, A.; Singh, S.; Von Kügelgen, J.; Hassani, H.; Pappas, G.J.; Schölkopf, B. Probable domain generalization via quantile risk minimization. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LO, USA & Online, 28 November–9 December; Volume 35, pp. 17340–17358.
15. Dubois, Y.; Ruan, Y.; Maddison, C.J. Optimal representations for covariate shifts. In Proceedings of the Thirty-Fifth Annual Conference on Neural Information Processing Systems (NeurIPS 2021), Online, 6–14 December 2021.
16. Blanchard, G.; Deshmukh, A.A.; Dogan, U.; Lee, G.; Scott, C. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.* **2021**, *22*, 1–55.
17. Dayal, A.; KB, V.; Cenkeramaddi, L.R.; Mohan, C.; Kumar, A.; Balasubramanian, N.V. MADG: Margin-based Adversarial Learning for Domain Generalization. In Proceedings of the Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 58938–58952.
18. Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Binas, J.; Zhang, D.; Le Priol, R.; Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 5815–5826.
19. Zhang, M.M.; Marklund, H.; Dhawan, N.; Gupta, A.; Levine, S.; Finn, C. Adaptive risk minimization: A meta-learning approach for tackling group shift. *arXiv* **2020**. [CrossRef]
20. Li, Y.; Yang, Y.; Zhou, W.; Hospedales, T. Feature-critic networks for heterogeneous domain generalization. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 3915–3924.
21. Li, D.; Yang, Y.; Song, Y.-Z.; Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
22. Shi, Y.; Seely, J.; Torr, P.H.; Siddharth, N.; Hannun, A.; Usunier, N.; Synnaeve, G. Gradient matching for domain generalization. *arXiv* **2021**. [CrossRef]
23. Rame, A.; Dancette, C.; Cord, M. Fishr: Invariant gradient variances for out-of-distribution generalization. In Proceedings of the International Conference on Machine Learning, Baltimore, MA, USA, 17–23 July 2022; pp. 18347–18377.
24. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. MixStyle Neural Networks for Domain Generalization and Adaptation. *Int. J. Comput. Vis.* **2024**, *132*, 822–836. [CrossRef]
25. Gulrajani, I.; Lopez-Paz, D. In search of lost domain generalization. *arXiv* **2020**. [CrossRef]
26. Foret, P.; Kleiner, A.; Mobahi, H.; Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv* **2020**. [CrossRef]
27. Du, J.; Yan, H.; Feng, J.; Zhou, J.T.; Zhen, L.; Goh, R.S.M.; Tan, V.Y. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv* **2021**. [CrossRef]

28. Liu, Y.; Mai, S.; Chen, X.; Hsieh, C.J.; You, Y. Towards Efficient and Scalable Sharpness-Aware Minimization. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12350–12360. [CrossRef]
29. Zhuang, J.; Gong, B.; Yuan, L.; Cui, Y.; Adam, H.; Dvornek, N.; Tatikonda, S.; Duncan, J.; Liu, T. Surrogate gap minimization improves sharpness-aware training. *arXiv* **2022**. [CrossRef]
30. Wang, P.; Zhang, Z.; Lei, Z.; Zhang, L. Sharpness-aware gradient matching for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3769–3778.
31. Wilson, G.; Cook, D.J. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 51. [CrossRef] [PubMed]
32. Zhao, S.; Yue, X.; Zhang, S.; Li, B.; Zhao, H.; Wu, B.; Krishna, R.; Gonzalez, J.E.; Sangiovanni-Vincentelli, A.L.; Seshia, S.A.; et al. A Review of Single-Source Deep Unsupervised Visual Domain Adaptation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 473–493. [CrossRef] [PubMed]
33. Zhang, Y.; Wei, Y.; Wu, Q.; Zhao, P.; Niu, S.; Huang, J.; Tan, M. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Trans. Image Process.* **2020**, *29*, 7834–7844. [CrossRef]
34. Li, R.; Jiao, Q.; Cao, W.; Wong, H.-S.; Wu, S. Model adaptation: Unsupervised domain adaptation without source data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9641–9650.
35. Liang, J.; Hu, D.; Feng, J. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In Proceedings of the International Conference on Machine Learning, Online, 12–18 July 2020; pp. 6028–6039.
36. Yang, S.; Wang, Y.; Van De Weijer, J.; Herranz, L.; Jui, S. Generalized source-free domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8978–8987.
37. Li, D.; Wu, A.; Wang, Y.; Han, Y. Prompt-Driven Dynamic Object-Centric Learning for Single Domain Generalization. *arXiv* **2024**. [CrossRef]
38. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain generalization with mixstyle. *arXiv* **2021**. [CrossRef]
39. Kim, D.; Yoo, Y.; Park, S.; Kim, J.; Lee, J. SelfReg: Self-supervised Contrastive Regularization for Domain Generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9599–9608. [CrossRef]
40. Nam, H.; Lee, H.; Park, J.; Yoon, W.; Yoo, D. Reducing Domain Gap by Reducing Style Bias. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8686–8695. [CrossRef]
41. Jiang, Y.; Neyshabur, B.; Mobahi, H.; Krishnan, D.; Bengio, S. Fantastic generalization measures and where to find them. *arXiv* **2019**. [CrossRef]
42. Zhang, X.; Xu, R.; Yu, H.; Zou, H.; Cui, P. Gradient norm aware minimization seeks first-order flatness and improves generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20247–20257.
43. Kaur, S.; Cohen, J.; Lipton, Z.C. On the maximum hessian eigenvalue and generalization. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LO, USA & Online, 28 November–9 December 2022; pp. 51–65.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.S.; Zhang, L. Bag-of-Visual-Words Scene Classifier With Local and Global Features for High Spatial Resolution Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. [CrossRef]
46. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
47. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
48. Zhao, L.; Tang, P.; Huo, L. Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *J. Appl. Remote Sens.* **2016**, *10*, 035004. [CrossRef]
49. Li, D.; Yang, Y.; Song, Y.Z.; Hospedales, T.M. Deeper, Broader and Artier Domain Generalization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5543–5551. [CrossRef]
50. Fang, C.; Xu, Y.; Rockmore, D.N. Unbiased Metric Learning: On the Utilization of Multiple Datasets and Web Images for Softening Bias. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1657–1664. [CrossRef]
51. Venkateswara, H.; Eusebio, J.; Chakraborty, S.; Panchanathan, S. Deep Hashing Network for Unsupervised Domain Adaptation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5385–5394. [CrossRef]
52. Chen, L.; Zhang, Y.; Song, Y.; Shan, Y.; Liu, L. Improved test-time adaptation for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 24172–24182.
53. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Kai, L.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
54. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]

55. Yan, S.; Song, H.; Li, N.; Zou, L.; Ren, L. Improve unsupervised domain adaptation with mixup training. *arXiv* **2020**. [CrossRef]
56. Li, H.; Pan, S.J.; Wang, S.; Kot, A.C. Domain Generalization with Adversarial Feature Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5400–5409. [CrossRef]
57. Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; Tao, D. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 647–663. [CrossRef]
58. Sagawa, S.; Koh, P.W.; Hashimoto, T.B.; Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv* **2019**. [CrossRef]
59. Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 11–14 October 2016; pp. 443–450. [CrossRef]
60. Pezeshki, M.; Kaba, O.; Bengio, Y.; Courville, A.C.; Precup, D.; Lajoie, G. Gradient starvation: A learning proclivity in neural networks. In Proceedings of the Thirty-Fifth Annual Conference on Neural Information Processing Systems (NeurIPS 2021), Online, 6–14 December 2021; Volume 34, pp. 1256–1272.
61. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
62. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
63. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features From Tiny Images*; University of Toronto: Toronto, ON, Canada, 2009.