EXPLORING MODE CONNECTIVITY IN KRYLOV SUBSPACE FOR DOMAIN GENERALIZATION

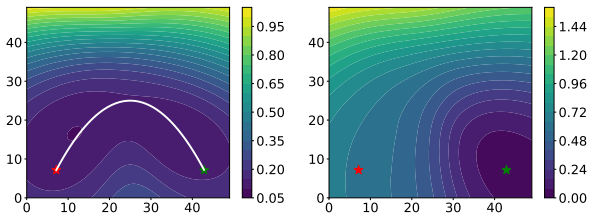
Anonymous authors
Paper under double-blind review

ABSTRACT

This paper explores the geometric characteristics of loss landscapes to enhance domain generalization (DG) in deep neural networks. Existing methods mainly leverage the local flatness around minima for improved generalization. However, recent theoretical studies indicate that flatness does not universally guarantee better generalization. Instead, this paper investigates a global geometrical property for domain generalization, i.e., *mode connectivity*, the phenomenon where distinct local minima are connected by continuous low-loss pathways. Different from flatness, mode connectivity enables transitions from poor to superior generalization models without leaving low-loss regions. To navigate these connected pathways effectively, this paper proposes a novel Billiard Optimization Algorithm (BOA), which discovers superior models by mimicking billiard dynamics. During this process, BOA operates within a low-dimensional Krylov subspace, aiming to alleviate the curse of dimensionality caused by the high-dimensional parameter space of deep models. Furthermore, this paper reveals that oracle test gradients strongly align with the Krylov subspace constructed from training gradients across diverse datasets and architectures. This alignment offers a powerful tool to bridge training and test domains, enabling the efficient discovery of superior models with limited training domains. Experiments on DomainBed demonstrate that BOA consistently outperforms existing sharpness-aware and DG methods across diverse datasets and architectures. Impressively, BOA even surpasses the sharpness-aware minimization by 3.6% on VLCS when using a ViT-B/16 backbone.

1 INTRODUCTION

Understanding the geometry of loss landscapes (Li et al., 2018a; Xu et al., 2024) in deep neural networks has emerged as a powerful and insightful approach for interpreting the generalization behavior of these models (Wu et al., 2017; Rangamani et al., 2020). Recent empirical and theoretical studies (Keskar et al., 2016; Dziugaite & Roy, 2017; Jiang et al., 2019) have demonstrated a strong connection between generalization performance and the geometric properties of the loss surface, particularly its sharpness. It has been observed that flat minima sought by sharpness-aware minimization (SAM) tend to generalize better than sharp ones (Foret et al., 2020; Kwon et al., 2021; Kim et al., 2022). This underscores the potential of leveraging insights from loss geometry.



(a) **Training Loss Landscape** (b) **Test Loss Landscape**

Figure 1: **Ideal and non-ideal models in the training and test loss landscapes.** (a) In the training loss landscape, the ideal (★) and non-ideal (★) models are connected via a low loss path. (b) The ideal model (★) is located in the basin of the test loss landscape, while the non-ideal model (★) is not.

Although numerous flatness-seeking methods have been developed to learn generalizable models (Kwon et al., 2021; Kim et al., 2022; Zhang et al., 2024), emerging research (Zhang et al., 2021; Wen et al., 2023) indicates that flatness does not universally guarantee improved generalization. This reveals limitations in over-relying on local sharpness measures, particularly within challenging Domain Generalization (DG) settings. In the DG task, our goal is to learn a model from multiple

Table 1: **Comparisons with other optimization algorithms in update subspaces and rules.** The update direction of SGD/SGLD lies in a random subspace $\mathcal{R}$; SAM lies in a two-dimensional Krylov subspace $\mathcal{K}_2$; and our BOA lies in a K-dimensional Krylov subspace $\mathcal{K}_K$.

Methods	Involving Subspaces	Update Rules
SGD/SGLD (Welling & Teh, 2011)	$\mathcal{R} = \{\boldsymbol{g}, \boldsymbol{z}\}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.	$\Delta\boldsymbol{\theta} = -\epsilon\boldsymbol{g} + \sqrt{2\epsilon\tau}\boldsymbol{z}$
SAM (Foret et al., 2020)	$\mathcal{K}_2 = \{\boldsymbol{g}, \mathbf{H}\boldsymbol{g}\}, \quad \mathbf{H}\boldsymbol{g} \approx \frac{\nabla\ell(\boldsymbol{\theta}+\epsilon\boldsymbol{g}) - \nabla\ell(\boldsymbol{\theta})}{\epsilon}$.	$\Delta\boldsymbol{\theta} = -\eta\nabla\ell(\boldsymbol{\theta} + \epsilon\boldsymbol{g})$
BOA (Ours)	$\mathcal{K}_K = \{\boldsymbol{g}, \mathbf{H}\boldsymbol{g}, \ldots, \mathbf{H}^{K-1}\boldsymbol{g}\}$.	$\begin{cases} \boldsymbol{p}_0 = -\sum_{k=0}^{K-1} \mathbf{H}^k \boldsymbol{g}, \\ \boldsymbol{p}_i = \left(\mathbf{I} - 2\frac{(\mathrm{proj}_{\mathcal{K}_K}\boldsymbol{g})(\mathrm{proj}_{\mathcal{K}_K}\boldsymbol{g})^\top}{\|\mathrm{proj}_{\mathcal{K}_K}\boldsymbol{g}\|^2}\right)\boldsymbol{p}_{i-1} \end{cases}$

source domains and generalize it to unseen target domains with different data distributions (Wang et al., 2022; Zhou et al., 2021; Shen et al., 2021). In such scenarios, the connection between flatness and generalization becomes significantly more complex and context-dependent. Crucially, the relationship between flatness in parameter space and the geometry of the representation space suggests that low sharpness may not adequately capture a model's vulnerability to domain shifts in feature space (Andriushchenko et al., 2023). Thus, although pursuing flat minima can be beneficial, it should not be considered a unique solution for DG. It is essential to investigate other geometric properties of the loss landscape beyond flatness to achieve better generalization.

In contrast to local flatness, this paper focuses on a global geometric property termed *mode connectivity* (Freeman & Bruna, 2016; Garipov et al., 2018; Draxler et al., 2018). That is, distinct local minima, discovered via independent training from different initializations, are connected by continuous pathways of low loss. This paper empirically investigates this phenomenon in the context of domain generalization (DG) and observes a similar connectivity: As visually illustrated in Figure 1, basins corresponding to models with significantly different out-of-domain performance (e.g., a non-ideal model with poor DG accuracy and an ideal model with nearly 100% DG accuracy) are connected via low-loss pathways. It suggests that the loss landscape is not composed of isolated basins but exhibits a connected structure among separate solutions. This connectivity implies a promising possibility: transitioning from a solution with poor generalization properties to one with strong out-of-domain performance without escaping the low-loss region. Despite its theoretical promise, effectively navigating these pathways in high-dimensional parameter spaces remains challenging due to the curse of dimensionality. Conventional optimization methods, such as SGD and its Langevin dynamics extensions (Bussi & Parrinello, 2007; Welling & Teh, 2011), are often ineffective in this context, as they tend to become trapped in local regions (Deng et al., 2020; Zheng et al., 2024), preventing the discovery of superior solutions. This problem motivates the need for a more deliberate algorithm capable of actively navigating these low-loss tunnels.

Inspired by the dynamics of billiard motion (Bunimovich, 2007; Gutkin, 2003), this paper proposes a novel and efficient traversal algorithm, Billiard Optimization Algorithm (BOA), which explicitly encourages movement along these connected pathways for improved domain generalization. Specifically, BOA operates through two core operations: (1) *line search* to locate loss contour boundaries (analogous to a billiard ball approaching a cushion), and (2) *reflection* to redirect the optimization trajectory upon boundary contact (mimicking momentum-preserving bounces). During this process of high-dimensional optimization, the curse of dimensionality profoundly impacts BOA's performance in two key ways. First, high-dimensional theory (Vershynin, 2018) suggests that random vectors become nearly orthogonal to the optimal initial directions (such as directions of oracle test gradients), making naive random direction search highly inefficient in locating useful paths. Second, trajectory sparsity emerges in high-dimensional landscapes, meaning that numerous optimization steps are required to achieve sufficient domain generalization. These issues are compounded in deep models, where the parameter space can be overwhelmingly large. Fortunately, a key geometric regularity identified in our empirical studies offers a solution: the test gradient exhibits strong alignment with the Krylov subspace (Liesen & Strakos, 2013) derived from training gradients across different datasets and architectures. The alignment effectively bridges the gap between training and test domains, and provides near-optimal initial directions for BOA without requiring access to test data. Furthermore, by leveraging this alignment, BOA can constrain its traversal trajectory to a reduced subspace of merely 5-20 dimensions, drastically reducing the vast search space and enabling efficient discovery of models with superior domain generalization capabilities. Empirical validation conducted on the challenging DomainBed (Gulrajani & Lopez-Paz, 2020) benchmark with various vision transformer architectures demonstrates that BOA consistently outperforms popular sharpness-aware methods and other prevalent DG techniques across five diverse datasets. These

results underscore the effectiveness of BOA in navigating high-dimensional loss landscapes and its practical utility for real-world domain generalization tasks.

Our key contributions can be summarized as follows:

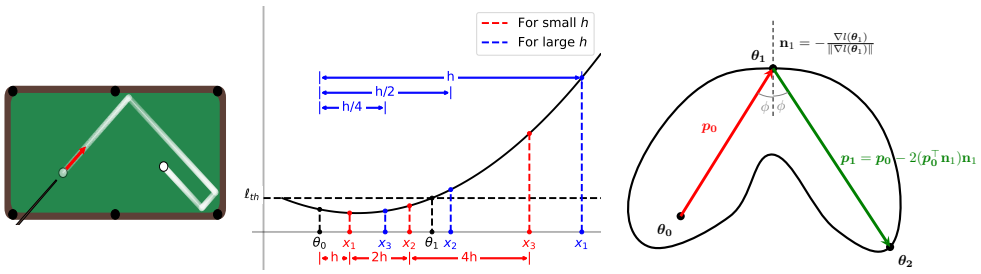
- Mode connectivity is identified in the DG context. It describes a phenomenon whereby continuous low-loss trajectories in the parameter space connect models exhibiting substantially divergent out-of-domain performance.
- A novel and efficient Billiards Optimization Algorithm (BOA) is introduced to advance domain generalization. It promotes navigation along low-loss paths connecting distinct local optima, facilitating the identification of models with enhanced generalization capabilities.
- Our study reveals the notable geometric regularity that test gradients demonstrate significant alignment with the Krylov subspace derived from training gradients, thereby establishing an effective bridge between training and unseen test domains.
- Experiments across diverse architectures confirm that BOA consistently surpasses popular sharpness-aware methods and other DG techniques on five datasets of DomainBed.

2 Related Work

Although deep learning has achieved success in many application areas, the loss landscapes of deep neural networks remain inadequately understood. This area constitutes an actively evolving field of research, primarily divided into two distinct categories.

Local structure. The first category explores the local structure of minima found by SGD and its variants. Researchers have observed that smaller mini-batch sizes often lead to sharp minima, while larger mini-batch sizes tend to yield flat minima (Keskar et al., 2016). In recent years, numerous studies have established a connection between flatness near minimizers and model generalization (Keskar et al., 2016; Dziugaite & Roy, 2017; Jiang et al., 2019): flat minima exhibit stronger generalization capabilities, whereas sharp minima perform poorly on test datasets. This conclusion has also been extended to out-of-distribution generalization scenarios (Zou et al., 2024), attracting widespread attention in the domain generalization research community and inspiring a series of domain generalization methods that seek flat minima, such as SWAD (Cha et al., 2021), SAM (Foret et al., 2020) and its variants (Wang et al., 2023; Li et al., 2025). Unfortunately, however, flatness does not equate to generalization. It is because: (1) There exist sharp yet well-generalizing models. Dinh *et al.* (Dinh et al., 2017) and Kaiyue Wen *et al.* (Wen et al., 2023) have theoretically and empirically identified minimizers that generalize well despite being very sharp. (2) Although some theoretical studies have demonstrated, under simple linear model settings, that the flattest neural network minimizers generalize well, this is not necessarily true for standard neural networks—even for simple architectures like two-layer ReLU networks. Unlike previous domain generalization methods, this paper leverages the connectivity of minimizers and attempts to traverse from any given minimizer to one with strong generalization capability by simulating billiard dynamics.

Global structure. The other major category of research focuses on the global structure of minima. Over the past few years, numerous studies have revealed connectivity between minima, known as mode connectivity. Freeman *et al.* (Freeman & Bruna, 2016) theoretically proved that local minima of a neural network with a single hidden layer and ReLU activation can be connected by a curve along which the loss value is bounded by a constant, which depends on the number of network parameters and the "smoothness" of the data. Their theoretical results are not easily generalizable to multi-layer networks. In contrast, Garipov *et al.* (Garipov et al., 2018) proposed a simpler training procedure that finds polygonal chains with nearly constant accuracy even in various modern state-of-the-art architectures, with only one bend between optima. Draxler *et al.* (Draxler et al., 2018) simultaneously and independently discovered curves connecting local optima in DNN loss landscapes. They adopted a different approach to finding these curves, inspired by the "Nudged Elastic Band" method from quantum chemistry. Currently, mode connectivity is used in algorithm design across multiple application areas, such as machine unlearning (Cheng & Amiri, 2025), model merging (Li et al., 2024), and continual learning (Mirzadeh et al., 2020). Although mode connectivity offers the potential to find models with stronger generalization capabilities continuously, this property has not been effectively utilized in the design of out-of-distribution generalization algorithms



(a) **Billiard Game** (b) **Line Search** (c) **Reflection**

Figure 2: **Overview of the proposed Billiard Optimization Algorithm (BOA).** (a) A schematic diagram of a billiards game on a two-dimensional table; (b) The first operation of BOA (Line Search): Like a ball approaching a cushion, this operation identifies loss contour boundaries through line search along directional vectors. (c) The second operation of BOA (Reflection): Upon boundary contact, the trajectory redirects via physics-based rules, mimicking a ball's reflection off a cushion.

due to the curse of dimensionality. To the best of our knowledge, this paper is the first work to seek models with better out-of-domain generalization via mode connectivity.

3 METHODOLOGY

Our *Billiard Optimization Algorithm* (BOA) aims to effectively navigate connected pathways within loss landscapes and discover models with superior out-of-domain generalization. Its design draws inspiration from the physical dynamics of billiards, as illustrated in Fig. 2a. Specifically, it iteratively performs two core operations: (1) *line search* to locate the loss contour boundary, analogous to a billiard ball moving toward a cushion (Fig. 2b), and (2) *reflection* to redirect the optimization trajectory upon reaching the boundary, mimicking a ball's reflection off a cushion (Fig. 2c). Due to the space limit, a formalized sketch describing the complete procedure of the Billiard Optimization Algorithm (BOA) is provided in Appendix A.

3.1 DEFINITION

In the Billiard Optimization Algorithm (BOA), the conceptual "billiard table" is mathematically defined as the sublevel set of the training loss landscape bounded by a specified contour threshold. Formally, this table corresponds to the parameter domain:

$$\mathcal{T} := \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \mid \ell_{\text{train}}(\boldsymbol{\theta}) \leq \ell_{\text{th}} \right\}, \tag{1}$$

where the contour threshold ℓ_{th} is deliberately constructed as: $\ell_{\text{th}} := \ell_{\text{train}}(\boldsymbol{\theta}_0) + \Delta_\ell$. Here, $\boldsymbol{\theta}_0$ represents the parameter vector of a pre-trained model, and $\Delta_\ell > 0$ denotes a strictly positive loss increment. This construction guarantees that the pre-trained model resides strictly within the interior of the billiard table, satisfying the containment condition: $\ell(\boldsymbol{\theta}_0) < \ell_{\text{th}}$. This foundational definition establishes the optimization landscape as a bounded playfield where subsequent billiard-inspired operations—contour-seeking line searches and momentum-preserving reflections—are performed. The bounded nature of the playfield effectively constrains variations in training loss, ensuring that performance on the training set remains nearly constant while the algorithm searches for model parameters that deliver superior DG performance.

3.2 LINE SEARCH

During the line search phase of the Billiard Optimization Algorithm (BOA), we emulate the linear trajectory of a billiard ball approaching a cushion by systematically locating the target loss contour boundary $\ell_{\text{th}} = \ell(\boldsymbol{\theta}_0) + \Delta_\ell$. Unless specified otherwise, we use ℓ to represent ℓ_{train} below.

This procedure mathematically corresponds to solving the nonlinear equation with respect to α:

$$\ell(\boldsymbol{\theta}_{i-1} + \alpha \boldsymbol{p}_{i-1}) = \ell_{\text{th}}, \tag{2}$$

Table 2: **Out-of-domain accuracies of ViT-B/16 on the DomainBed benchmark.** The star symbol ($\star$) marks experiments involving full fine-tuning of the image encoder, with its absence signifying the application of visual prompt tuning instead.

Algorithms	VLCS	PACS	OfficeHome	TerraIncognita	DomainNet	Avg.
ERM* (Vapnik, 1998)	81.6	93.2	80.0	54.5	57.7	73.4
MIRO* (Cha et al., 2022)	83.6	95.8	82.3	58.8	57.2	75.5
ERM (Jia et al., 2022)	81.9	95.9	84.1	56.1	59.5	75.5
IRM (Arjovsky et al., 2019)	82.9	96.1	83.2	56.7	59.1	75.6
DANN (Ganin et al., 2016)	81.8	96.3	83.0	56.0	58.4	75.1
CDANN (Li et al., 2018c)	82.4	96.5	82.9	55.6	58.4	75.2
MMD (Li et al., 2018b)	82.3	95.8	83.6	57.4	<u>59.9</u>	75.8
CORAL (Sun & Saenko, 2016)	82.6	96.4	83.8	57.5	59.8	76.0
IIB (Li et al., 2022b)	82.3	96.5	84.2	58.2	58.6	76.0
CSVPT (Li et al., 2022a)	82.9	96.6	85.6	<u>58.6</u>	59.1	76.5
SAM (Foret et al., 2020)	82.9	96.6	85.4	56.2	59.8	76.2
GSAM (Zhuang et al., 2022)	82.9	96.6	<u>85.6</u>	55.4	59.8	76.1
GAM (Zhang et al., 2023)	<u>83.6</u>	96.4	85.5	55.3	59.5	76.1
SAGM (Wang et al., 2023)	82.8	96.8	85.2	58.0	59.1	76.4
DISAM (Zhang et al., 2024)	82.7	<u>97.1</u>	85.4	57.3	59.8	<u>76.5</u>
BOA (Ours)	**86.5**	**97.4**	**86.0**	**60.3**	**60.2**	**78.1**

where $\boldsymbol{\theta}_{i-1}$ represents the initial parameter vector (analogous to the billiard ball's starting position) and $\boldsymbol{p}_{i-1}$ denotes the search direction. Candidate points are generated along the ray $\boldsymbol{x}_k = \boldsymbol{\theta}_{i-1} + h_k \boldsymbol{p}_{i-1}$ ($h_k > 0$), with BOA employing an adaptive bracketing strategy to efficiently isolate the solution interval $[h_L, h_R]$: when $\ell(\boldsymbol{x}_k) < \ell_{\text{th}}$, step sizes expand exponentially via $h_{k+1} = (2^k - 1)h$ (where h represents the initial step) until exceeding ℓ_{th}; conversely, when $\ell(\boldsymbol{x}_k) > \ell_{\text{th}}$, step sizes contract geometrically through $h_k = h_{k-1}/2$ until falling below ℓ_{th}. Following interval isolation, BOA employs the golden-section search method for precise refinement:

$$h_k^{(1)} = h_R - \frac{h_R - h_L}{\psi},\ h_k^{(2)} = h_L + \frac{h_R - h_L}{\psi}, \tag{3}$$

where $\psi = (1 + \sqrt{5})/2$ represents the golden ratio. The interval boundaries $[h_L, h_R]$ are then updated based on function evaluations at points $h_k^{(1,2)}$. Finally, the approximate optimal solution $\alpha^\star$ of equation (2), found via the golden-section search, is used to update the parameters:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} + \alpha^\star \boldsymbol{p}_{i-1}, \tag{4}$$

precisely positioning the model parameter on the loss contour.

3.3 Reflection

Upon reaching the contour boundary, BOA initiates its second operation: *reflection* that mimics momentum-preserving billiard collisions.

At the contour point $\boldsymbol{\theta}_i$, the reflection direction is computed by leveraging the geometry of the local loss landscape, analogous to the physical reflection process illustrated in Fig. 2c. In this phase, the unit normal vector is expressed as: $\boldsymbol{n}_i = -\frac{\nabla \ell(\boldsymbol{\theta}_i)}{\|\nabla \ell(\boldsymbol{\theta}_i)\|_2}$, which corresponds to the steepest descent direction, plays a role equivalent to the surface normal of the cushion in billiard dynamics. As depicted in Fig. 2c, both the incident direction $\boldsymbol{p}_{i-1}$ and the normal vector $\boldsymbol{n}_i$ define the reflection plane. The specular reflection direction is subsequently updated as follows:

$$\boldsymbol{p}_i = \boldsymbol{p}_{i-1} - 2(\boldsymbol{p}_{i-1}^\top \boldsymbol{n}_i)\boldsymbol{n}_i = (\mathbf{I} - 2\boldsymbol{n}_i \boldsymbol{n}_i^\top)\boldsymbol{p}_{i-1}. \tag{5}$$

Geometrically, this transformation subtracts twice the orthogonal projection of $\boldsymbol{p}_{i-1}$ onto $\boldsymbol{n}_i$. This operation preserves the magnitude of momentum, i.e., $\|\boldsymbol{p}_i\|_2 = \|\boldsymbol{p}_{i-1}\|_2$, and adheres to the reflection law $\phi_{\text{incident}} = \phi_{\text{reflected}}$, which implies $(\boldsymbol{p}_{i-1}, \boldsymbol{n}_i) = (\boldsymbol{p}_i, \boldsymbol{n}_i)$. The reflection operator $R[\boldsymbol{n_i}] = \mathbf{I} - 2\boldsymbol{n}_i \boldsymbol{n}_i^\top$ represents an improper rotation (with $\det R = -1$) that incorporates curvature information via $\boldsymbol{n}_i \propto \nabla \ell(\boldsymbol{\theta}_i)$. This allows efficient exploration along the loss contour without

the need for Hessian recomputation, thereby maintaining a computational complexity of $\mathcal{O}(d)$ per iteration for d-dimensional parameters.

After multiple iterations of alternating line search and reflection operations, we generate a trajectory of model parameters and select an optimal model from this trajectory via a validation set for testing.

3.4 Overcoming the "Curse of Dimensionality"

The curse of dimensionality fundamentally impacts optimization in high-dimensional parameter spaces through two primary mechanisms: (1) In billiard dynamics, the initial incident direction $\boldsymbol{p}_0$ determines the trajectory's ability to find model parameters with better DG performance. Unlike intuitive 2D billiard, the high-dimensional setting induces a geometric constraint: random vectors become nearly orthogonal to the oracle optimal directions (Vershynin, 2018), suggesting that naive random search becomes extremely inefficient in high dimensions. (2) Trajectory sparsity emerges in high-dimensional landscapes, necessitating numerous optimization steps to achieve sufficient out-of-domain generalization.

To overcome the challenges of high-dimensional optimization, we leverage the geometric structure of the Krylov subspace, which is generated from training derivatives:

$$\mathcal{K}_K(\mathbf{H}_{\text{train}}, \boldsymbol{g}_{\text{train}}) = \text{span}\left\{\boldsymbol{g}_0,\ \mathbf{H}\boldsymbol{g}_0,\ \ldots,\ \mathbf{H}^{K-1}\boldsymbol{g}_0\right\}, \tag{6}$$

where $\boldsymbol{g}_0 = \nabla\ell_{\text{train}}(\boldsymbol{\theta}_0)$ and $\mathbf{H} = \nabla^2\ell_{\text{train}}(\boldsymbol{\theta}_0)$. Krylov subspace (Liesen & Strakos, 2013) is famous as an efficient, low-dimensional framework to find approximate solutions to high-dimensional linear algebra problems. Here, we use the Krylov subspace concerning the Hessian matrix to capture the dominant curvature information. It provides two critical advantages: First, it enables principled selection of the initial search direction through efficient approximation of test gradients. Second, it constrains the trajectory to a subspace with limited dimensions that already includes model parameters with better generalization.

Determine the initial incident direction. Intuitively, in the DG scenario, the test gradient $\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0)$ might serve as a good choice for the initial incident direction. Fortunately, empirical analysis reveals a remarkable geometric regularity: the test gradient $\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0)$ exhibits strong alignment with the Krylov subspace $\mathcal{K}_K(\mathbf{H}_{\text{train}}, \boldsymbol{g}_{\text{train}})$. This alignment enables efficient approximation via setting proper values for β_k:

$$\boldsymbol{p}_0 = \sum_{k=0}^{K-1} \beta_k \mathbf{H}^k \boldsymbol{g}_0 \approx \nabla\ell_{\text{test}}(\boldsymbol{\theta}_0). \tag{7}$$

Interestingly, experimental evidence demonstrates that setting $\beta_k = -1$ yields a particularly effective direction. This configuration achieves a small angle between $\boldsymbol{p}_0$ and $\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0)$ across benchmark datasets, effectively overcoming dimensional barriers. To avoid explicit Hessian matrix calculation, we employ the finite difference to approximate Hessian-vector products:

$$\mathbf{H}\boldsymbol{g}_0 \approx \frac{\nabla\ell(\boldsymbol{\theta}_0 + \epsilon\boldsymbol{g}_0) - \nabla\ell(\boldsymbol{\theta}_0)}{\epsilon}, \tag{8}$$

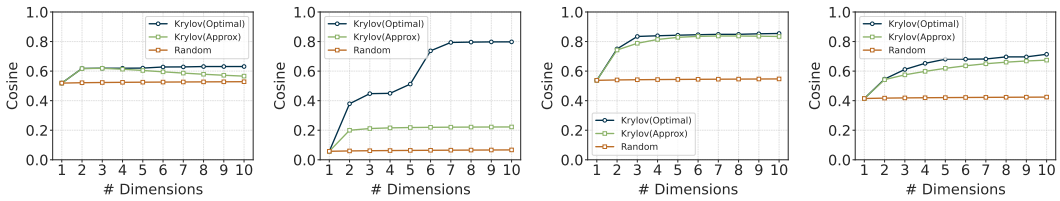
maintaining $\mathcal{O}(d)$ complexity while circumventing second-order derivative calculations.

Constrains the trajectory to limited dimensions. The Krylov alignment mentioned above also means the existence of better OOD solutions within the Krylov subspaces. To limit the reflection operation within the Krylov subspace, the reflection direction $\boldsymbol{p}_i$ can be updated as follows:

$$\boldsymbol{p}_i = (\mathbf{I} - 2\tilde{\boldsymbol{n}}_i\tilde{\boldsymbol{n}}_i^\top)\boldsymbol{p}_{i-1}, \tag{9}$$

where $\tilde{\boldsymbol{n}}_i = \text{proj}_{\mathcal{K}_K}\boldsymbol{n}_i$ denotes the projection of $\boldsymbol{n}_i$ within the Krylov subspace $\mathcal{K}_K$.

The Krylov alignment constitutes a "free lunch" in high-dimensional optimization, providing near-optimal initial directions without the availability of test data and reducing the search space substantially. This geometric regularity bridges the gap between training and test distributions, enabling effective exploration of loss landscapes for improved domain generalization.



(a) Caltech101 (b) LabelMe (c) SUN09 (d) VOC2007

Figure 3: **Cosine similarity between test gradients and their approximations derived from either random or Krylov subspaces.** Krylov(Optimal): This curve corresponds to the projection of the test gradients onto the Krylov subspace, representing the best possible approximation achievable under the given basis; Krylov(Approx): This curve corresponds to an approximation obtained via the heuristic choice of $\beta_k = -1$ in equation (7); Random: This curve depicts the theoretical expectation value of the projection length of the unit test gradients onto random subspaces (see Appendix B).

4 Experiments

4.1 Experimental setting

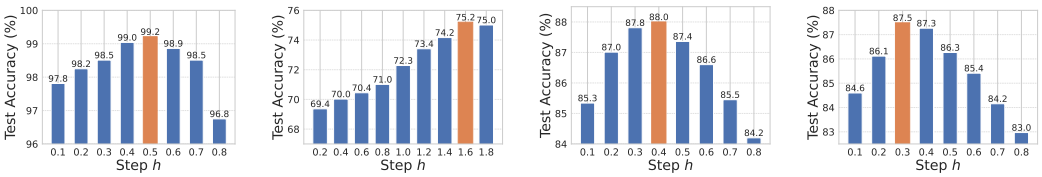
Datasets and Protocol. Our evaluation is conducted on five widely-used, challenging domain generalization datasets: PACS (Li et al., 2017) (9,991 images, 4 domains and 7 classes), VLCS (Fang et al., 2013) (10,729 images, 4 domains and 5 classes), OfficeHome (Venkateswara et al., 2017) (15,588 images, 4 domains and 65 classes), TerraIncognita (Beery et al., 2018) (24,788 images, 4 domains and 10 classes), and DomainNet (Peng et al., 2019) (586,575 images, 6 domains and 345 classes). Following the standardized DomainBed benchmark (Gulrajani & Lopez-Paz, 2020), we implement a rigorous leave-one-domain-out evaluation protocol: each experiment designates one domain as the test set (i.e., test or target domain) while aggregating all remaining domains for training (i.e., training or source domain). To ensure robust model selection and hyperparameter tuning, previous research typically reserves 20% of the data from each source domain to create a validation set. However, in our experiments, models with similar validation accuracy can exhibit over 10% difference in DG performance. Thus, relying on this set amounts to random guessing, leading to the use of a test-domain validation set as a necessary compromise. For a fair comparison with prior DG methods, we re-evaluate all baselines using the same test-domain validation set.

Implementation Details. Given the demonstrated parameter efficiency and effectiveness of Visual Prompt Tuning (VPT) in domain generalization benchmarks like DomainBed (Jia et al., 2022; Li et al., 2022a; Zheng et al., 2022), our study primarily focuses on VPT-based optimization. To validate the broad applicability of our approach, we conduct comprehensive experiments across three backbones: ViT-B/32, ViT-B/16, and ViT-L/14. Our framework builds upon SAM-trained pretrained models, with specific configurations: we use batch sizes of 16 for ViT-B/32 and ViT-B/16, and 8 for ViT-L/14 to accommodate GPU memory constraints, while maintaining a fixed learning rate of 5e-4 across all architectures. All experiments employ the Adam optimizer for consistent optimization. During the billiard optimization, we fix the step h at 10, and set the reflection time to 20 iterations. Other hyperparameters, including the Krylov subspace dimension K and loss variation Δ_ℓ, are determined through grid search, with model selection using the validation set.

4.2 Krylov Subspace Analysis

Prior to the formal evaluation of our BOA algorithm, we present a comprehensive analysis of the Krylov subspace introduced to overcome the curse of dimensionality in high-dimensional optimization.

Firstly, to validate the feasibility of the billiard dynamics, we conduct experiments with oracle test domain: when using the ideal test gradient $\nabla \ell_{\text{test}}(\boldsymbol{\theta}_0)$ as the incident direction, BOA achieves significant performance gains across five datasets (Table 4). The ERM+BOA and SAM+BOA configurations demonstrate consistent improvements averaging +4.9% and +5.0%, respectively, over baseline methods, confirming the feasibility of the billiard dynamics and the ideality of the test gradient as the initial incident direction.



(a) Caltech101 (b) LabelMe (c) SUN09 (d) VOC2007

Figure 4: OOD accuracies along approximate initial incident directions on four VLCS domains.

Table 3: Out-of-domain test accuracies with different ViT backbones.

Algorithms	VLCS	PACS	OH	TI	DN	Avg.
Backbone: ViT-B/32						
ERM	82.4	94.5	79.6	36.6	54.3	69.5
SAM	81.9	95.1	80.6	42.1	54.6	70.9
BOA (Ours)	**84.9**	**96.3**	**80.9**	**49.7**	**54.9**	**73.3**
Backbone: ViT-B/16						
ERM	81.9	95.9	84.1	56.1	59.5	75.5
SAM	82.9	96.6	85.4	56.2	59.8	76.2
BOA (Ours)	**86.5**	**97.4**	**86.0**	**60.3**	**60.2**	**78.1**
Backbone: ViT-L/14						
ERM	82.7	98.6	90.2	61.3	64.6	79.5
SAM	82.5	98.2	90.9	64.1	64.8	80.1
BOA (Ours)	**86.4**	**98.7**	**91.2**	**65.7**	**65.4**	**81.5**

Table 4: Test accuracies when using the oracle test gradients as the initial incident directions.

Algorithms	VLCS	PACS	OH	TI	DN	Avg.
ERM	81.9	95.9	84.1	56.1	59.5	75.5
ERM+**BOA (Ours)**	**85.6**	**97.7**	**86.3**	**69.5**	**62.9**	**80.4**
SAM	82.9	96.6	85.4	56.2	59.8	76.2
SAM+**BOA (Ours)**	**87.0**	**97.9**	**87.6**	**69.6**	**62.9**	**81.2**

Table 5: Comparisons of out-of-domain accuracy when using a ResNet50 architecture.

Method	Caltech101	LabelMe	SUN09	VOC2007	Avg.
ERM	98.2	67.1	73.0	77.6	79.0
SAM	99.6	65.7	75.1	80.5	80.2
BOA	**99.7**	**69.0**	**77.2**	**81.1**	**81.8**

Fortunately, we discovered the test gradient exhibits strong alignment with the training-derived Krylov subspace, enabling effective approximation in the absence of test data. Specifically, two facts can be observed in Figure 5: (1) Rapid improvement in directional alignment (quantified by $\cos\gamma_K$) as subspace dimension K increases can be observed, where $\cos\gamma_K$ is defined by the cosine similarity between test gradient and its optimal approximation with the Krylov subspace $\mathcal{K}_K$:

$$\cos\gamma_K = \max_{\boldsymbol{p}\in\mathcal{K}_K} \cos\langle\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0), \boldsymbol{p}\rangle = \cos\langle\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0), \text{proj}_{\mathcal{K}_K}\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0)\rangle, \tag{10}$$

where $\langle\cdot,\cdot\rangle$ denotes the angle between two vectors. Remarkably, $\cos\gamma_K$ exceeds 0.8 for $K = 7 \sim 10$ on "LabelMe"—an eightfold improvement over the $K = 1$ case ($\cos\gamma_1 < 0.1$). This sharply contrasts with random subspaces, where alignment increases marginally at best. More detailed analysis about Krylov and random subspace can be found in Appendix B. (2) The approximate initial incident direction defined in equation (7) has a similar increasing trend with K, achieving near-optimal alignment on "SUN09" and "VOC2007". Note that the approximate initial incident direction is obtained without using test data, different from the optimal approximation that utilizes $\text{proj}_{\mathcal{K}_K}\nabla\ell_{\text{test}}(\boldsymbol{\theta}_0)$. Additionally, we even find a significant increase in test accuracies along the approximate initial incident direction (+1.4% on "Caltech101", +6.5% on "LabelMe", +2.7% on "SUN09" and +2.9% on "VOC2007", see Figure 4). However, this performance gain occurs concurrently with training loss degradation, necessitating BOA's reflection mechanism to maintain the training loss as much as possible while navigating toward generalizable solutions.

4.3 Main Results

Comparisons with SOTA domain generalization methods. Table 2 provides a decisive performance comparison of domain generalization methods on the DomainBed benchmark, clearly establishing our BOA method as the new state-of-the-art. With an average accuracy of 78.1% across all five datasets, BOA outperforms the previous best methods (DISAM at 76.5%) by a significant 1.6% margin. This performance delta represents the largest inter-method improvement observed in the benchmark, demonstrating BOA's exceptional cross-domain generalization capabilities. The dataset-specific analysis reveals particularly striking results; On VLCS, BOA achieves a remarkable 86.5% accuracy—surpassing SAM's 82.9% by a dramatic 3.6% margin. Notably, BOA achieves these record results using only visual prompt tuning rather than full encoder fine-tuning, making its efficiency particularly impressive. The consistent leadership across all datasets—from relatively

easy domains like PACS to exceptionally difficult ones like TerraIncognita—demonstrates BOA's ability to navigate varied domain shifts.

Experiments with different Vision Transformers. Based on the comprehensive experimental results presented in Table 3, the proposed BOA method demonstrates consistent and superior effectiveness across all evaluated Vision Transformer backbones—namely ViT-B/32, ViT-B/16, and ViT-L/14—achieving the highest average accuracy on the multi-domain benchmarks compared to both ERM and SAM. Under the ViT-B/32 backbone, BOA attains an average of 73.3%, outperforming SAM (70.9%) and ERM (69.5%) by margins of 2.4% and 3.8%, respectively, with particularly notable gains on challenging domains such as TI (49.7% vs SAM's 42.1%). This advantage is maintained with larger backbones: using ViT-B/16, BOA reaches 78.1% on average, surpassing SAM by 1.9% and ERM by 3.6%, and with ViT-L/14, it achieves 81.5%, exceeding SAM by 1.4% and ERM by 2.0%. The performance improvement is consistent not only in the aggregate metric but across all five individual datasets, underscoring BOA's robustness and its capability to enhance generalization irrespective of the model capacity or architectural variant of the backbone network.

Broad applicability to CNN architectures. While much of our analysis has focused on Vision Transformer architectures, we now demonstrate that the core findings can extend effectively to standard CNN backbones. To this end, we conduct additional experiments using a ResNet50 backbone to validate the persistence of Krylov alignment in CNN architectures. As summarized in Table 6, the superiority of Krylov subspaces over random subspaces remains consistent across multiple datasets, similar to the trends observed in ViT architectures (as illustrated in Figure 3). Notably, on the SUN09 dataset, Krylov alignment (quantified by $\cos\gamma_K$) exceeds 0.75 for $K = 10$, marking an improvement of approximately +0.5 over the $K = 1$ case. This suggests that Krylov subspaces capture geometrically meaningful directions in the loss landscape, even in CNNs.

Table 6: **Cosine similarity between test gradients and their approximations derived from either random or Krylov subspaces (using a ResNet50 backbone).**

Domain	Subspace Type	Subspace Dimension (K)				
		2	4	6	8	10
Caltech101	Random	0.3121	0.3122	0.3124	0.3125	0.3125
	Krylov	**0.3317**	**0.3339**	**0.3360**	**0.3381**	**0.3381**
LabelMe	Random	0.4063	0.4064	0.4065	0.4066	0.4067
	Krylov	**0.4937**	**0.5235**	**0.5724**	**0.5825**	**0.5855**
SUN09	Random	0.2653	0.2654	0.2655	0.2657	0.2657
	Krylov	**0.4199**	**0.7547**	**0.7603**	**0.7610**	**0.7617**
VOC2007	Random	0.1395	0.1396	0.1397	0.1399	0.1399
	Krylov	**0.1406**	**0.2991**	**0.3260**	**0.3414**	**0.3707**

We also evaluate the DG performance of our BOA method using a ResNet50 backbone. As illustrated in Table 5, BOA achieves the highest accuracy across all four domains, with an average accuracy of 81.8%, outperforming both ERM (79.0%) and SAM (80.2%). The consistency in performance gains reinforces BOA's utility as an architecture-agnostic optimization framework. Notably, in our ResNet experiments, we observed that most hyperparameters could be directly transferred from ViT-B/16 setups. However, due to differences in loss landscape geometry (e.g., CNNs may exhibit sharper minima), we found that maintaining the step size $h = 10$ requires excessive line search iterations to identify solution intervals. To ensure efficient optimization, we adjusted the step size to $h = 1$ for ResNet50.

4.4 Ablation Study

Figure 5 presents a decisive ablation study on one VLCS domain ("LabelMe"), evaluating the contribution of Krylov subspace components to the Billiard Optimization Algorithm (BOA) across three loss variation thresholds ($\Delta_\ell = 0.05, 0.1, 0.3$). In Figure 5, Krylov-inspired $\boldsymbol{p}_0$ outperforms random initialization by 0.5-1.6% across all Δ_ℓ values, confirming the effectiveness of Krylov subspace on the determination of $\boldsymbol{p}_0$. Across all three Δ_ℓ values, the complete BOA within Krylov subspace method consistently outperforms intermediate implementations. At $\Delta_\ell = 0.3$, it achieves a remarkable 75.7% accuracy—6.5% higher than random initialization (69.2%) and 4.9% superior to Krylov-

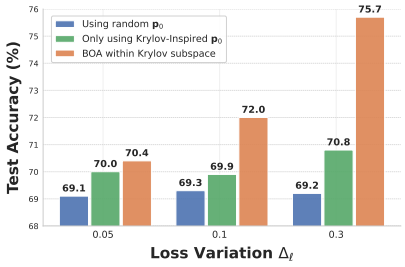
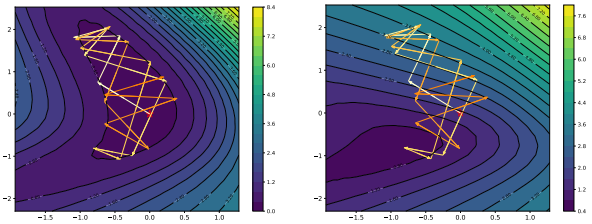


Figure 5: Ablation study on the effect of Krylov subspace using one VLCS domain ("LabelMe") for test.



(a) **Training Loss Landscape** (b) **Test Loss Landscape**

Figure 6: **Trajectory Visualization of BOA.** The same trajectory, derived from the training loss landscape, is shown in both the (a) training and (b) test loss landscapes.

inspired direction alone (70.8%). This substantial margin demonstrates that Krylov-enhanced reflection operations provide compounded benefits beyond initial direction selection.

4.5 Loss Landscape Visualization

Our loss landscape visualization examines two key aspects: (1) connectivity between models with different out-of-domain performance, and (2) optimization trajectory analysis of the Billiard Optimization Algorithm (BOA).

Mode Connectivity Analysis. A baseline model was first trained on three OfficeHome domains ("Clipart", "Product", and "Real World"), which may exhibit poor performance when generalized to the unseen "Art" domain. In contrast, an ideal Oracle model was trained using all domains. Subsequently, building upon the approach in (Garipov et al., 2018), we connect these two models using a parameterized Bézier curve and optimize it to maintain uniformly low loss throughout the entire pathway. Finally, as illustrated in Figure 1, a 2D visualization demonstrates a continuous low-loss valley bridging the non-ideal (red marker) and ideal (green marker) models. Remarkably, this connectivity pattern has been consistently observed across diverse datasets and architectures.

BOA Trajectory Visualization. To visually demonstrate the effectiveness of the BOA algorithm, we employed constrained Krylov subspaces with a dimension of $K = 2$ to trace its optimization path on the VLCS dataset. As illustrated in Figure 6, BOA successfully navigates the training loss landscape and seeks out-of-domain solutions located within optimal regions on the test loss surface. This visualization offers clear visual evidence of BOA's capability in discovering better models.

5 Conclusion

This study demonstrates that leveraging mode connectivity in loss landscapes offers a novel and effective approach for enhancing domain generalization. We empirically established that low-loss pathways connect models with divergent generalization capabilities, revealing a geometric structure that enables transitions to superior solutions without leaving low-loss regions. To navigate this structure, we proposed the Billiard Optimization Algorithm (BOA), which efficiently traverses these connected modes using physics-inspired operations. A key insight enabling this efficiency was the strong alignment between test gradients and the low-dimensional Krylov subspace derived from training gradients, allowing BOA to operate effectively in a reduced subspace. Integrated with parameter-efficient vision prompt tuning, BOA achieved superior performance across multiple DomainBed benchmarks, consistently outperforming existing sharpness-aware and domain generalization techniques. These results highlight the value of geometric connectivity properties for developing robust models that generalize well to unseen domains. Future work could explore extending BOA to other geometric properties of loss landscapes or adapting it for dynamic environments where domain shifts occur continuously.

References

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36: 47032–47051, 2023.

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

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Leonid Bunimovich. Dynamical billiards. *Scholarpedia*, 2(8):1813, 2007.

Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 75(5):056707, 2007.

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

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *arXiv preprint arXiv:2203.10789*, 2022.

Jiali Cheng and Hadi Amiri. Understanding machine unlearning through the lens of mode connectivity. *arXiv preprint arXiv:2504.06407*, 2025.

Wei Deng, Guang Lin, and Faming Liang. A contour stochastic gradient langevin dynamics algorithm for simulations of multi-modal distributions. *Advances in neural information processing systems*, 33:15725–15736, 2020.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Eugene A Gutkin. Billiard dynamics: a survey with the emphasis on open problems on billiards. *Regular and chaotic dynamics*, 8(1):1–13, 2003.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Aodi Li, Liansheng Zhuang, Shuo Fan, and Shafei Wang. Learning common and specific visual prompts for domain generalization. In *Proceedings of the Asian Conference on Computer Vision*, pp. 4260–4275, 2022a.

Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, and Shafei Wang. Seeking consistent flat minima for better domain generalization via refining loss landscapes. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 15349–15359, June 2025.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7399–7407, 2022b.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018b.

Wenyi Li, Huan-ang Gao, Mingju Gao, Beiwen Tian, Rong Zhi, and Hao Zhao. Training-free model merging for multi-target domain adaptation. In *European Conference on Computer Vision*, pp. 419–438. Springer, 2024.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, 2018c.

Jörg Liesen and Zdenek Strakos. *Krylov subspace methods: principles and analysis*. Numerical Mathematics and Scie, 2013.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. *arXiv preprint arXiv:2010.04495*, 2020.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Akshay Rangamani et al. *Loss landscapes and generalization in neural networks: Theory and applications*. PhD thesis, Johns Hopkins University, 2020.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

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

Vladimir Vapnik. Statistical learning theory. *(No Title)*, 1998.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3769–3778, 2023.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36:1024–1035, 2023.

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Yichu Xu, Xin-Chun Li, Lan Li, and De-Chuan Zhan. Visualizing, rethinking, and mining the loss landscape of deep neural networks. *arXiv preprint arXiv:2405.12493*, 2024.

Ruipeng Zhang, Ziqing Fan, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Domain-inspired sharpness-aware minimization under domain shifts. *arXiv preprint arXiv:2405.18861*, 2024.

Shuofeng Zhang, Isaac Reid, Guillermo Valle Pérez, and Ard Louis. Why flatness does and does not correlate with generalization for deep neural networks. *arXiv preprint arXiv:2103.06219*, 2021.

Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, and Peng Cui. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20247–20257, 2023.

Haoyang Zheng, Hengrong Du, Qi Feng, Wei Deng, and Guang Lin. Constrained exploration via reflected replica exchange stochastic gradient langevin dynamics. *arXiv preprint arXiv:2405.07839*, 2024.

Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*, 2022.

Yingtian Zou, Kenji Kawaguchi, Yingnan Liu, Jiashuo Liu, Mong-Li Lee, and Wynne Hsu. Towards robust out-of-distribution generalization bounds via sharpness. *arXiv preprint arXiv:2403.06392*, 2024.