Beyond Average Flatness: Domain-wise Flatness for Domain Generalization

Seungjun Choi **Heeyoung Kim***
Department of Industrial and Systems Engineering, KAIST
{seungjun@kaist.ac.kr, heeyoungkim@kaist.ac.kr}

Abstract

Sharpness-Aware Minimization (SAM) is widely used in domain generalization to promote flat minima that generalize to unseen distributions. Existing SAM-based methods typically penalize the sharpness of the *average* loss across domains. We show that this prevailing formulation—termed AvSAM—has a fundamental geometric limitation: it can favor solutions with misaligned domain-wise dominant Hessian eigenvectors. As a result, such solutions may appear flat on average while remaining sharp within individual domains. To address this limitation, we propose an alternative objective that penalizes the average *domain-wise* sharpness. We show that this objective decomposes into the AvSAM objective plus a non-negative residual term, which we call the *Cancellation Gap*. This gap is minimized if and only if the domain-wise dominant eigenvectors are aligned up to sign, indicating that it mitigates AvSAM's tendency to favor misaligned solutions. However, we further uncover a degeneracy: the Cancellation Gap can become small without improving alignment when sharpness is highly imbalanced across domains. To mitigate this failure mode, we propose Cancellation Gap-Aware Minimization (CGM), which augments the objective with a squared coefficient-of-variation (CV^2) regularizer to discourage sharpness imbalance. Experiments show that CGM achieves the best average performance across standard domain generalization benchmarks among strong baselines. Diagnostic analyses further confirm reduced domain-wise sharpness and improved alignment of dominant Hessian eigenvectors.

1 Introduction

Modern deep neural networks achieve impressive performance across a wide range of tasks, yet their accuracy often degrades when the test distribution deviates from the training domains. Domain generalization (DG) seeks to address this challenge by learning models that generalize to unseen domains without access to target-domain data. Prior work has explored a variety of approaches, including representation alignment [1–3], meta-learning [4], data augmentation [5–7], and regularization-based methods [8–11]. Recent studies suggest that solutions lying in flat regions of the loss landscape tend to generalize better [12–15], motivating sharpness-aware training approaches. In particular, Sharpness-Aware Minimization (SAM) [16] improves robustness by minimizing the worst-case loss within a small neighborhood in parameter space. As a result, incorporating SAM into DG frameworks has emerged as a promising direction. Within this growing line of research, the prevailing paradigm applies SAM to the *average* loss across source domains [17–20], a strategy we refer to as AvSAM. Equivalently, AvSAM optimizes the average loss together with its sharpness, implicitly assuming that flatness of the average loss is sufficient for domain generalization.

While effective, AvSAM suffers from a fundamental geometric limitation: it may favor solutions whose domain-wise dominant Hessian eigenvectors are misaligned. As illustrated in Figure 1,

*corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

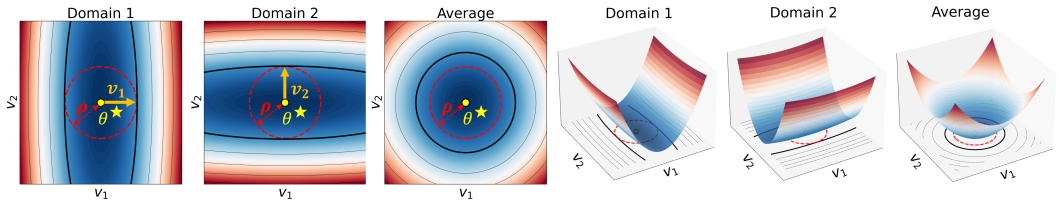


Figure 1: Illustration of sharpness cancellation under misaligned domain-wise dominant Hessian eigenvectors. Around a joint local minimizer $\theta^\star$, the dominant eigenvectors of two source domains are orthogonal ($v_1 \perp v_2$). Although each domain loss is sharp within the ρ-neighborhood along its dominant direction, averaging the losses can *cancel* these domain-specific worst-case increases, making the average loss appear relatively flat. Black bold contours indicate equal loss levels.

misalignment can lead to sharpness cancellation under averaging, making the average loss appear flat even though domain-wise losses remain sharp. Such solutions lack uniform robustness across source domains, making them undesirable for DG. Our analysis further shows that stronger alignment among domain-wise dominant Hessian eigenvectors induces a larger curvature-based lower bound on average-loss sharpness, providing a geometric explanation for AvSAM's tendency to favor solutions with weaker alignment.

To address this limitation, we introduce Individual-domain SAM (InSAM), an alternative objective that replaces average-loss sharpness with the average of domain-wise sharpness. Intuitively, penalizing sharpness for each domain individually should yield more uniformly robust solutions than penalizing only the average-loss sharpness. We show that the InSAM objective decomposes into the AvSAM objective plus a residual term:

$$\mathcal{L}_{\mathrm{InSAM}}(\theta) = \mathcal{L}_{\mathrm{AvSAM}}(\theta) + \underbrace{\left(\frac{1}{D} \sum_{i=1}^{D} \mathcal{S}_i(\theta) - \mathcal{S}_{\mathrm{Av}}(\theta) \right)}_{\text{Cancellation Gap}}.$$

The residual, termed the *Cancellation Gap*, measures the discrepancy between average domain-wise sharpness and average-loss sharpness, capturing sharpness masked by averaging (here, θ denotes model parameters, D the number of domains, $\mathcal{S}_i(\theta)$ the sharpness of domain i, and $\mathcal{S}_{\mathrm{Av}}(\theta)$ the sharpness of the average loss). We show that this gap is non-negative and is minimized if and only if the domain-wise dominant Hessian eigenvectors are aligned up to sign. Thus, InSAM can be viewed as AvSAM regularized by the Cancellation Gap, discouraging solutions that are flat-on-average yet sharp-per-domain and counteracting AvSAM's tendency to favor weakly aligned solutions.

However, naively minimizing this objective introduces a potential pitfall. We identify a *bad shortcut*: the Cancellation Gap can be reduced not through improved directional alignment, but by allowing the domain-wise sharpness distribution to be dominated by a single domain. To prevent this imbalance-driven degeneracy, we propose **Cancellation Gap-Aware Minimization (CGM)**, which augments InSAM with a squared coefficient-of-variation (CV^2) regularizer over domain-wise sharpness. This regularizer discourages sharpness imbalance and encourages the gap to be reduced through improved alignment rather than trivial dominance. As a result, CGM better realizes the intended goal of InSAM: reducing average domain-wise sharpness while avoiding flat-on-average yet sharp-per-domain solutions. Our contributions are summarized as follows:

- We identify a fundamental geometric limitation of AvSAM arising from weak alignment among domain-wise dominant Hessian eigenvectors, and show how InSAM addresses it via the Cancellation Gap.
- We uncover a bad shortcut: the Cancellation Gap can be reduced by increasing sharpness imbalance across domains rather than by improving directional alignment.
- We propose CGM, which mitigates this issue using a CV^2 regularizer. CGM achieves the best average performance across benchmarks among strong DG baselines, with diagnostic analyses showing reduced domain-wise sharpness and improved alignment.

2 Preliminaries and related works

2.1 Preliminaries

Domain generalization (DG). Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space. We consider D distinct source domains indexed by $\mathcal{I} = \{1, \ldots, D\}$. Each domain $i \in \mathcal{I}$ corresponds to a joint distribution P_i over $\mathcal{X} \times \mathcal{Y}$. These source domains differ in style, environment, acquisition conditions, or other factors. From each P_i, we observe a dataset $\mathcal{D}_i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a predictor parametrized by $\theta \in \Theta$. The goal of DG is to learn f_θ from source datasets $\{\mathcal{D}_i\}_{i=1}^{D}$ so that it generalizes to an unseen target distribution $P_T \neq P_i$. For $i \in \mathcal{I}$, the empirical loss on domain i is $\mathcal{L}_i(\theta) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(f_\theta(x_j^i), y_j^i)$, where ℓ is a pointwise loss (e.g., cross-entropy). Following standard DG practice, we optimize the average training loss $\mathcal{L}_{\mathrm{Av}}(\theta) = \frac{1}{D} \sum_{i \in \mathcal{I}} \mathcal{L}_i(\theta)$.

Sharpness-aware minimization (SAM). Sharpness-Aware Minimization (SAM) [16] is an optimization framework aimed at finding flat minima. Given a loss $\mathcal{L}(\theta)$, SAM defines the sharpness $\mathcal{S}^\rho(\theta) := \max_{\|\varepsilon\|_2 \leq \rho} (\mathcal{L}(\theta + \varepsilon) - \mathcal{L}(\theta))$ as the maximum increase in loss within an ℓ_2-ball of radius $\rho > 0$ centered at θ. SAM then minimizes the sum of the loss and the sharpness, which is equivalent to solving $\max_{\|\varepsilon\|_2 \leq \rho} \mathcal{L}(\theta + \varepsilon)$. In practice, SAM uses the first-order Taylor approximation of this objective: $\mathcal{L}(\theta + \tilde{\varepsilon}) \approx \max_{\|\varepsilon\|_2 \leq \rho} \mathcal{L}(\theta + \varepsilon)$, where $\tilde{\varepsilon} = \rho \frac{\nabla \mathcal{L}(\theta)}{\|\nabla \mathcal{L}(\theta)\|_2}$. Differentiating the approximated objective with respect to θ introduces higher-order terms due to the dependence of $\tilde{\varepsilon}$ on θ. In practice, SAM adopts a first-order approximation that ignores these terms, and updates the parameters using the gradient evaluated at the perturbed point, $\nabla \mathcal{L}(\theta + \tilde{\varepsilon})$.

2.2 Related works

Risk-based robustness and domain balancing. GroupDRO [21] minimizes the worst-group risk, while V-REx [11] penalizes the variance of risks across training domains. Although CGM also addresses cross-domain imbalance, it differs in both the quantity being regularized and the underlying purpose. Rather than balancing domain risks, CGM penalizes the dispersion of domain-wise sharpness. Importantly, this penalty serves as a safeguard against degenerate reductions of the Cancellation Gap driven by sharpness dominance, rather than enforcing equal sharpness across domains. We provide a detailed objective- and gradient-level comparison in Appendix B.

Inconsistency-aware DG methods. Beyond risk-level robustness, another line of DG research addresses cross-domain loss geometry inconsistency. Fish [8] encourages agreement among domain-wise gradients, AND-mask [22] masks gradient coordinates with inconsistent signs, and recent methods such as Arith [23] and GGA [24] reduce gradient conflicts by steering optimization toward more balanced or aligned update directions. Another direction considers higher-order inconsistency: Fishr [9] matches domain-level gradient variances to match domain-wise risks and Hessian structure near optima. Our work is similarly motivated by cross-domain geometric inconsistency, but focuses on a specific curvature-level inconsistency: weak alignment among domain-wise dominant Hessian eigenvectors. Beyond treating such inconsistency as undesirable, we provide a sharpness-based explanation for why it is harmful for DG.

SAM and flatness-based DG methods. A growing body of work has linked flatter minima to improved generalization [12–15]. While this connection was studied primarily in the i.i.d. setting, SWAD [25] extended flatness-based ideas to DG by constructing flat solutions via trajectory averaging. Unlike explicit sharpness regularization, however, trajectory averaging does not directly penalize sharpness in the training objective. SAM [16] explicitly targets parameter-space sharpness, and several related methods have explored flatness-based optimization for DG. GAM [26] introduces first-order flatness, while FAD [18] jointly considers zeroth- and first-order flatness for DG. SAGM [17] designs a perturbation strategy to implicitly align the perturbed and standard gradients of the average loss, while UDIM [19] combines parameter- and input-space perturbations to reduce loss-landscape inconsistency between source and perturbed domains. DISAM [20] introduces perturbations informed by domain-wise loss variance to mitigate convergence inconsistency, and GAC-FAS [27] aligns domain-wise perturbed gradients with ERM gradients for face anti-spoofing DG. More recently, SFT [28] examines cross-domain loss-landscape inconsistency under SAM, characterizing it through discrepancies in domain-wise sharpness values. However, SFT takes a different approach from optimizer-level SAM variants: it introduces an additional self-feedback framework and refines the loss landscape using a soft-label generator, rather than directly modifying the sharpness penalty or SAM update rule.

3 Geometric analysis of sharpness in DG

In this section, we analyze the geometry of AvSAM in DG. Unless otherwise specified, we adopt the standard notion of sharpness from the original SAM literature, defined as the maximum loss increase within an ℓ_2 ball of radius $\rho > 0$. For domain $i \in \mathcal{I}$ and the average loss $\mathcal{L}_{\mathrm{Av}}(\theta)$, we define the domain-wise sharpness and average-loss sharpness as

$$\mathcal{S}_i^\rho(\theta) := \max_{\|\varepsilon\|_2 \le \rho} \left(\mathcal{L}_i(\theta+\varepsilon) - \mathcal{L}_i(\theta)\right), \qquad \mathcal{S}_{\mathrm{Av}}^\rho(\theta) := \max_{\|\varepsilon\|_2 \le \rho} \left(\mathcal{L}_{\mathrm{Av}}(\theta+\varepsilon) - \mathcal{L}_{\mathrm{Av}}(\theta)\right). \quad (1)$$

In non-convex DG, joint solutions are generally non-unique. We assume the existence of (possibly multiple) joint local minimizers shared across all source domains, and denote their set by

$$\Theta^\star := \{\theta \in \Theta : \ \theta \text{ is a local minimizer of } \mathcal{L}_i \text{ for all } i \in \mathcal{I}\}.$$

Following prior DG loss-landscape analyses [9, 22, 29], we study the local geometry around these joint minimizers. For the local Hessian-based analysis, we assume that each $\mathcal{L}_i$ is twice continuously differentiable in a neighborhood of $\theta \in \Theta^\star$. We denote $H_i(\theta) := \nabla^2 \mathcal{L}_i(\theta)$ and $H_{\mathrm{Av}}(\theta) := \frac{1}{D}\sum_{i\in\mathcal{I}} H_i(\theta)$. The local optimality conditions then give $\nabla \mathcal{L}_i(\theta) = 0$ and $H_i(\theta) \succeq 0$ for all $\theta \in \Theta^\star$ and $i \in \mathcal{I}$.

3.1 AvSAM and its geometric limitation

Recall that AvSAM applies SAM to the average loss, thereby penalizing average-loss sharpness. Using the definitions in Eq. (1), the objective of AvSAM is

$$\mathcal{L}_{\mathrm{AvSAM}}(\theta) := \max_{\|\varepsilon\|_2 \le \rho} \mathcal{L}_{\mathrm{Av}}(\theta+\varepsilon) = \mathcal{L}_{\mathrm{Av}}(\theta) + \mathcal{S}_{\mathrm{Av}}^\rho(\theta).$$

The following proposition shows that the average-loss sharpness $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$ admits a curvature-based lower bound governed by the cross-domain alignment of domain-wise dominant Hessian eigenvectors. To measure this alignment in a sign-invariant manner, we define the spectral alignment score $\mathcal{A}_{\mathrm{spec}}(\theta) := \lambda_{\max}\big(\frac{1}{D}\sum_{i\in\mathcal{I}} v_i v_i^\top\big)$, where v_i is a unit dominant eigenvector of $H_i(\theta)$ and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. This score lies in $[1/D,\ 1]$, attaining 1 when all v_i are aligned up to sign, and $1/D$ when they are mutually orthogonal.

Proposition 3.1. ***(Spectral-Alignment Lower Bound for AvSAM)*** *For any $\theta \in \Theta^\star$, let v_i be a unit dominant eigenvector of the positive semidefinite Hessian $H_i(\theta)$ for each domain $i \in \mathcal{I}$, and define $\lambda_- := \min_{i\in\mathcal{I}} \lambda_{\max}(H_i(\theta))$. Then, for sufficiently small $\rho > 0$,*

$$\mathcal{S}_{\mathrm{Av}}^\rho(\theta) \approx \frac{\rho^2}{2}\lambda_{\max}(H_{\mathrm{Av}}(\theta)), \quad \textit{where} \quad \lambda_{\max}(H_{\mathrm{Av}}(\theta)) \ge \lambda_- \, \mathcal{A}_{\mathrm{spec}}(\theta).$$

The proof is provided in Appendix A.1. Proposition 3.1 shows that stronger spectral alignment induces a larger curvature-based lower bound on the average-loss sharpness $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$. Since AvSAM penalizes this term, candidates with higher alignment face a larger barrier to achieving small average-loss sharpness. Thus, when $\Theta^\star$ contains multiple joint minimizers with comparable empirical loss and curvature scales, AvSAM can be interpreted as implicitly favoring solutions with lower alignment. This bias toward misalignment is problematic because, as illustrated in Figure 1, it can induce sharpness cancellation under averaging, yielding *flat-on-average yet sharp-per-domain* minima.

3.2 InSAM and Cancellation Gap

Section 3.1 shows that penalizing average-loss sharpness $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$ can implicitly favor solutions with lower spectral alignment, potentially leading to *flat-on-average yet sharp-per-domain* minima. To instead promote flatness across all source domains, a natural alternative is to replace the average-loss sharpness with the average of domain-wise sharpness. We refer to this objective as Individual-domain SAM (InSAM). Using the definitions in Eq. (1), the objective of InSAM is

$$\mathcal{L}_{\mathrm{InSAM}}(\theta) := \frac{1}{D}\sum_{i\in\mathcal{I}} \max_{\|\varepsilon_i\|_2 \le \rho} \mathcal{L}_i(\theta+\varepsilon_i) = \mathcal{L}_{\mathrm{Av}}(\theta) + \frac{1}{D}\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta).$$

To quantify how InSAM differs from AvSAM, we introduce the *Cancellation Gap*.

Definition 3.2. (Cancellation Gap) The Cancellation Gap $\mathcal{C}_\rho(\theta)$ is defined as the difference between the average domain-wise sharpness and the average-loss sharpness:

$$\mathcal{C}_\rho(\theta) := \frac{1}{D}\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) - \mathcal{S}_{\mathrm{Av}}^\rho(\theta).$$

The Cancellation Gap quantifies the extent to which a solution may appear flat on average while remaining sharp on individual domains due to sharpness cancellation under averaging. By Definition 3.2, the InSAM objective decomposes exactly as

$$\mathcal{L}_{\mathrm{InSAM}}(\theta) = \mathcal{L}_{\mathrm{AvSAM}}(\theta) + \mathcal{C}_\rho(\theta).$$

Thus, InSAM can be interpreted as AvSAM regularized by $\mathcal{C}_\rho(\theta)$. The following theorem establishes its non-negativity in general and characterizes when it vanishes under a local quadratic model.

Assumption 3.3. (Local Quadratic Loss Model) For any $\theta \in \Theta^\star$ and each domain $i \in \mathcal{I}$, suppose that $H_i(\theta) \succeq 0$ and that the following holds for all $\|\varepsilon\|_2 \le \rho$:

$$\mathcal{L}_i(\theta+\varepsilon) - \mathcal{L}_i(\theta) = \frac{1}{2}\varepsilon^\top H_i(\theta)\varepsilon$$

Theorem 3.4. ***(Cancellation Gap and Spectral Alignment)***

1. ***Non-negativity.*** *For any $\theta \in \Theta$, we have $\mathcal{C}_\rho(\theta) \ge 0$.*
2. ***Zero-gap characterization (quadratic case).*** *Under Assumption 3.3, for any $\theta \in \Theta^\star$, suppose that the dominant eigenvalue of each $H_i(\theta)$ is positive and simple. Then*
$$\mathcal{C}_\rho(\theta) = 0 \iff \mathcal{A}_{\mathrm{spec}}(\theta) = 1.$$

The proof is provided in Appendix A.2. Theorem 3.4 establishes that the Cancellation Gap $\mathcal{C}_\rho(\theta)$ is always non-negative and, under the local quadratic model, equals zero if and only if spectral alignment is maximal. Therefore, in the local quadratic regime, $\mathcal{C}_\rho(\theta)$ can be interpreted as an alignment regularizer. From this perspective, augmenting the AvSAM objective with $\mathcal{C}_\rho(\theta)$ counteracts AvSAM's tendency to favor misaligned solutions, thereby helping InSAM to effectively mitigate the misalignment effects induced by AvSAM.

3.3 When Cancellation Gap fails: bad shortcut

We show that the Cancellation Gap $\mathcal{C}_\rho(\theta)$ can be small even under poor spectral alignment when the domain-wise sharpness is highly imbalanced and dominated by a single domain. We refer to this degenerate mechanism as a *bad shortcut*. To quantify sharpness imbalance, define the maximum domain-wise sharpness $\mathcal{S}_{\max}^\rho(\theta) := \max_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta)$ and the sharpness dominance ratio

$$w_{\max}(\theta) := \frac{\mathcal{S}_{\max}^\rho(\theta)}{\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta)} \in \left[\frac{1}{D}, 1\right].$$

We focus on the non-trivial case $\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) > 0$, since otherwise $\mathcal{C}_\rho(\theta) = 0$ holds trivially.

Theorem 3.5. ***(Bad Shortcut via Sharpness Dominance)*** *Under Assumption 3.3, for any $\theta \in \Theta^\star$ satisfying $\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) > 0$, the Cancellation Gap is upper-bounded as*

$$\mathcal{C}_\rho(\theta) \le \frac{1}{D}\Big(\sum\nolimits_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta)\Big)\big(1 - w_{\max}(\theta)\big).$$

The proof is provided in Appendix A.3. Theorem 3.5 reveals a failure mode of using the Cancellation Gap as an alignment regularizer: $\mathcal{C}_\rho(\theta)$ can be reduced without improving spectral alignment by making the domain-wise sharpness distribution highly skewed. In the extreme case where one domain dominates (i.e., $w_{\max}(\theta) \to 1$), $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$ becomes effectively governed by that domain. As a result, $\frac{1}{D}\sum_i \mathcal{S}_i^\rho(\theta)$ and $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$ nearly coincide, causing $\mathcal{C}_\rho(\theta)$ to be small even under poor alignment. This highlights the need for an additional mechanism that discourages sharpness imbalance, ensuring that reductions in $\mathcal{C}_\rho(\theta)$ are not achieved via such a degenerate shortcut.

4 Cancellation Gap-Aware Minimization (CGM): blocking the bad shortcut

To block the bad shortcut, we propose CGM, which retains the InSAM objective (and thus the Cancellation Gap) while explicitly penalizing *sharpness imbalance* across domains. Specifically, we regularize the dispersion of domain-wise sharpness via the squared coefficient-of-variation, CV^2, ensuring that reductions in the Cancellation Gap are not achieved through this degenerate mechanism.

4.1 CV^2 regularization for balanced sharpness

We define the squared coefficient-of-variation of domain-wise sharpness as

$$\mathrm{CV}^2(\theta) := \frac{\sigma^2(\theta)}{\mu^2(\theta)}, \quad \text{where} \quad \mu(\theta) := \frac{1}{D}\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta), \quad \sigma^2(\theta) := \frac{1}{D}\sum_{i\in\mathcal{I}} \left(\mathcal{S}_i^\rho(\theta) - \mu(\theta)\right)^2. \tag{2}$$

We focus on the non-trivial case $\mu(\theta) > 0$. The squared coefficient-of-variation $\mathrm{CV}^2(\theta)$ is a scale-invariant measure of dispersion, with $\mathrm{CV}^2(\theta) = 0$ if and only if $\mathcal{S}_1^\rho(\theta) = \cdots = \mathcal{S}_D^\rho(\theta)$, i.e., domain-wise sharpness is perfectly balanced. The following proposition formalizes that $\mathrm{CV}^2(\theta)$ upper-bounds sharpness dominance.

Proposition 4.1. ***($\mathrm{CV}^2(\theta)$ upper-bounds sharpness dominance)*** *Assume $\mu(\theta) > 0$. Let $\mathrm{CV}^2(\theta)$ be defined as in Eq. (2). Then,*

$$w_{\max}(\theta) \;\le\; \frac{1}{D}\left(1 + \sqrt{(D-1)\,\mathrm{CV}^2(\theta)}\right).$$

The proof is provided in Appendix A.4. Proposition 4.1 shows that regularizing $\mathrm{CV}^2(\theta)$ controls sharpness dominance, thereby blocking the bad shortcut in which the Cancellation Gap becomes small due to sharpness imbalance. We incorporate $\mathrm{CV}^2(\theta)$ as an additional regularizer on top of InSAM, yielding

$$\mathcal{L}_{\mathrm{CGM}}(\theta) := \mathcal{L}_{\mathrm{Av}}(\theta) + \frac{1}{D}\sum\nolimits_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) + \alpha\,\mathrm{CV}^2(\theta), \tag{3}$$

where $\alpha \ge 0$ controls the strength of the imbalance penalty; when $\alpha = 0$, CGM reduces to InSAM.

4.2 Optimization and algorithm

Eq. (3) defines the idealized CGM objective. In practice, however, $\mathrm{CV}^2(\theta)$ can be reduced either by decreasing the dispersion $\sigma^2(\theta)$ or by increasing the mean $\mu(\theta)$, the latter conflicting with the goal of promoting flatness across domains. To avoid this, we optimize a surrogate in which gradients through $\mu^2(\theta)$ are stopped, treating it as constant during backpropagation. Using SAM-based sharpness estimates $\widehat{\mathcal{S}}_i^\rho(\theta) := \mathcal{L}_i(\theta + \tilde{\varepsilon}_i) - \mathcal{L}_i(\theta)$ and their mean $\widehat{\mu}(\theta) := \frac{1}{D}\sum_{i\in\mathcal{I}} \widehat{\mathcal{S}}_i^\rho(\theta)$, the resulting gradient under the standard SAM first-order approximation is

$$\begin{aligned} \nabla\mathcal{L}_{\mathrm{CGM}}(\theta) &\approx \frac{1}{D}\sum\nolimits_{i\in\mathcal{I}} \Big((1-c_i)\nabla\mathcal{L}_i(\theta) + c_i\nabla\mathcal{L}_i(\theta+\tilde{\varepsilon}_i)\Big), \\ \text{where} \quad \tilde{\varepsilon}_i &= \rho\frac{\nabla\mathcal{L}_i(\theta)}{\|\nabla\mathcal{L}_i(\theta)\|_2} \quad \text{and} \quad c_i = 1 + \frac{2\alpha}{\widehat{\mu}^2(\theta)}\Big(\widehat{\mathcal{S}}_i^\rho(\theta) - \widehat{\mu}(\theta)\Big). \end{aligned} \tag{4}$$

A full derivation is provided in Appendix A.5. Importantly, the coefficient c_i increases with the estimated centered sharpness $\widehat{\mathcal{S}}_i^\rho(\theta) - \widehat{\mu}(\theta)$: sharper-than-average domains place larger weights on the perturbed (SAM) gradient $\nabla\mathcal{L}_i(\theta + \tilde{\varepsilon}_i)$, while flatter domains rely more on the standard gradient $\nabla\mathcal{L}_i(\theta)$. Since c_i is not guaranteed to lie in $[0, 1]$, the update can become extrapolative rather than interpolative between $\nabla\mathcal{L}_i(\theta)$ and $\nabla\mathcal{L}_i(\theta + \tilde{\varepsilon}_i)$. To stabilize optimization, we use the clipped coefficient

$$\bar{c}_i := \mathrm{clip}(c_i, 0, 1) = \min\{1, \max\{0, c_i\}\},$$

and replace c_i with $\bar{c}_i$ in practice. This yields an interpolative update: sharper domains typically retain full SAM weight, whereas flatter domains receive reduced weight. The clipped update should be viewed as a practical stabilization heuristic rather than the exact gradient of the surrogate. Further analyses of the clipped coefficient dynamics $\bar{c}_i$ and the effect of coefficient clipping are provided in Appendix C.2.

Algorithm 1 CGM (training procedure)

Require: Source domains $\mathcal{I} := \{1, \ldots, D\}$, per-domain minibatch size b, perturbation radius ρ, initial parameters θ, CV^2 penalty coefficient α, stabilizer ξ, base optimizer $\mathcal{O}$.

while not converged **do**
  for $i \in \mathcal{I}$ **do**
    Sample a minibatch $\mathcal{B}_i = \{(x_j^i, y_j^i)\}_{j=1}^b$.
    Compute minibatch loss $\mathcal{L}_i(\theta)$ and gradient $g_i^{(0)} = \nabla \mathcal{L}_i(\theta)$.
    Compute SAM perturbation $\tilde{\varepsilon}_i = \rho g_i^{(0)} / \|g_i^{(0)}\|_2$.
    Compute perturbed loss $\mathcal{L}_i(\theta + \tilde{\varepsilon}_i)$ and gradient $g_i^{(1)} = \nabla \mathcal{L}_i(\theta + \tilde{\varepsilon}_i)$.
    Compute sharpness estimate $\widehat{\mathcal{S}}_i^{\rho}(\theta) = \mathcal{L}_i(\theta + \tilde{\varepsilon}_i) - \mathcal{L}_i(\theta)$.
  end for
  Compute $\widehat{\mu}(\theta) = \frac{1}{D} \sum_{i \in \mathcal{I}} \widehat{\mathcal{S}}_i^{\rho}(\theta)$.
  Compute $c_i = 1 + 2\alpha(\widehat{\mathcal{S}}_i^{\rho}(\theta) - \widehat{\mu}(\theta)) / \left(\widehat{\mu}^2(\theta) + \xi\right)$, $\bar{c}_i = \mathrm{clip}(c_i, 0, 1)$ for all $i \in \mathcal{I}$.
  Compute $g = \frac{1}{D} \sum_{i \in \mathcal{I}} [(1 - \bar{c}_i) g_i^{(0)} + \bar{c}_i g_i^{(1)}]$.
  Update parameters: $\theta \leftarrow \mathcal{O}(\theta, g)$
end while

5 Experiments

5.1 Experimental setups

Datasets and baselines. We evaluate our method on five standard DG benchmarks from DomainBed [30]: PACS [31] (9,991 images, 7 classes, and 4 domains), VLCS [32] (10,729 images, 5 classes, and 4 domains), OfficeHome [33] (15,588 images, 65 classes, and 4 domains), TerraIncognita [34] (24,788 images, 10 classes, and 4 domains), and DomainNet [35] (586,575 images, 345 classes, and 6 domains). We compare CGM against two families of baselines: (i) widely used DG methods, including those addressing loss-geometry inconsistency [8, 9, 11, 21–24], and (ii) SAM-based methods that introduce no additional networks or modules [17, 18, 20, 26, 36].

Evaluation protocol. We use a standard leave-one-domain-out evaluation scheme: for each split, one domain is held out as the target domain for testing, while the remaining domains are used as source domains for training and validation. Model selection is based on source-domain validation, and performance is evaluated on the held-out target domain. The final accuracy is obtained by averaging over all target domains and three independent runs with different random seeds.

Implementation details. Following DomainBed, all methods implemented by us use an ImageNet [37]-pretrained ResNet-50 [38] as backbone and Adam [39] as the base optimizer. The CGM penalty coefficient α is sampled from $10^{\mathrm{Uniform}(-3.5,-1.5)}$, and the perturbation radius ρ from $\{0.01, 0.03, 0.05\}$. To mitigate noisy early domain-wise sharpness estimates, we apply an InSAM warm-up before switching to CGM; the warm-up ratio $\gamma \in \{0.1, 0.2, 0.3\}$ denotes the fraction of iterations using InSAM. Following recent baselines [17, 20, 23, 24, 36], we tune the remaining non-method-specific hyperparameters (batch size, learning rate, weight decay, and dropout rate) using the search protocol of SWAD [25]. Additional details, including the complete hyperparameter search space and GPU configurations, are provided in Appendix C.1. The project repository is available at https://github.com/SeungjunChoi1/CGM.

5.2 Main results

Table 1 reports average target-domain accuracy on five DG benchmarks. CGM achieves the best overall performance (66.4%) and attains the best or second-best result on four datasets (PACS, VLCS, OfficeHome and DomainNet). Compared with strong SAM-based baselines, CGM improves over SAGM by +0.3 points on average (66.1→66.4) and DISAM by +0.5 on average (65.9→66.4). Since SAGM achieves the second-best overall performance, we further compare SAGM and CGM under identical experimental settings and random seeds. The paired per-seed performance differences are reported in Appendix C.3. InSAM consistently outperforms AvSAM across all benchmarks (+0.7 average, 65.1→65.8), supporting our claim that average-loss flatness alone is insufficient for uniform cross-domain robustness. CGM further improves over InSAM on all five benchmarks (+0.6 on average, 65.8→66.4), highlighting the importance of controlling sharpness imbalance beyond penalizing average domain-wise sharpness.

Table 1: We report the average accuracy over leave-one-domain-out splits (mean ± standard error over 3 seeds). Results marked with $*$ are obtained from our own experiments, while those marked with † and ‡ are taken from [17] and [18], respectively. Other baseline results are taken from the original papers. The extended table including additional baselines is provided in Appendix C.5.

Category	Method	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg.
Non-SAM DG baselines	GroupDRO†	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
	V-REx†	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
	AND-mask†	84.4±0.9	78.1±0.9	65.6±0.4	44.6±0.3	37.2±0.6	62.0
	ERM†	85.5±0.2	77.3±0.4	66.5±0.3	46.1±1.8	43.8±0.1	63.9
	Fish†	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
	Fishr	85.5±0.4	77.8±0.1	67.8±0.1	47.4±1.6	41.7±0.0	64.0
	GGA	86.4±0.5	78.7±0.8	67.0±0.3	48.5±1.1	44.4±0.2	65.0
	Arith	86.5±0.3	79.4±0.8	69.4±0.3	48.1±1.2	41.5±0.2	65.0
SAM-based DG baselines	GSAM†	85.9±0.1	79.1±0.2	69.3±0.0	47.0±0.8	44.6±0.2	65.1
	GAM‡	86.1±0.6	78.5±0.4	68.2±1.0	45.2±0.6	43.8±0.1	64.4
	FAD‡	**88.2**±0.5	78.9±0.8	69.2±0.5	45.7±1.0	44.4±0.1	65.3
	SAGM†	86.6±0.2	80.0±0.3	70.1±0.4	**48.8**±0.9	45.0±0.2	66.1
	DISAM	87.1±0.2	79.9±0.5	70.3±0.2	46.6±0.8	45.4±0.2	65.9
	AvSAM*	86.5±0.4	78.9±0.3	69.5±0.1	45.9±0.9	44.8±0.1	65.1
	InSAM*	87.1±0.3	79.7±0.2	70.0±0.2	47.1±0.9	45.2±0.1	65.8
	CGM*	87.5±0.5	**80.1**±0.1	**70.5**±0.1	48.3±0.9	**45.6**±0.0	**66.4**

Table 2: Diagnostic metrics on PACS. We report average domain-wise sharpness, average-loss sharpness, Cancellation Gap, sharpness dominance ratio, and spectral alignment, averaged over leave-one-domain-out splits and five independent runs (mean ± standard error).

Method	$\frac{1}{D}\sum_i \widehat{\mathcal{S}}_i^\rho$	$\widehat{\mathcal{S}}_{\mathrm{Av}}^\rho$	$\widehat{\mathcal{C}}_\rho$	$\widehat{w}_{\max}$	$\mathcal{A}_{\mathrm{spec}}$
AvSAM	0.040±0.006	0.020±0.005	0.020±0.002	0.617±0.029	0.366±0.003
InSAM	0.033±0.005	0.021±0.003	0.012±0.006	0.713±0.046	0.378±0.006
CGM	0.030±0.003	0.019±0.003	0.011±0.005	0.619±0.030	0.412±0.007

5.3 Diagnostic analyses

Beyond target accuracy, we validate the geometric analysis on the PACS dataset using five source-domain diagnostics: average domain-wise sharpness $\frac{1}{D}\sum_i \mathcal{S}_i^\rho$, average-loss sharpness $\mathcal{S}_{\mathrm{Av}}^\rho$, the Cancellation Gap $\mathcal{C}_\rho$, sharpness dominance ratio $w_{\max}$, and spectral alignment $\mathcal{A}_{\mathrm{spec}}$. The first four are estimated via SAM-based loss increases. For spectral alignment, we approximate the domain-wise dominant Hessian eigenvectors via power iteration with Hessian-vector products. For a controlled comparison, we use a learning rate of 3×10^{-5}, dropout rate of 0.5, weight decay of 10^{-4}, and perturbation radius $\rho = 0.03$ across all methods. For CGM, we set $\alpha = 10^{-3}$ and $\gamma = 0.1$. For each leave-one-domain-out split, we compute diagnostics on source domains at the selected checkpoint, average over splits, and repeat across five independent runs. Results are summarized in Table 2.

AvSAM's sharpness cancellation and InSAM's sharpness imbalance. AvSAM exhibits a large gap between average-loss sharpness (0.020) and average domain-wise sharpness (0.040), yielding the largest Cancellation Gap (0.020). This is consistent with our analysis that AvSAM can induce large sharpness cancellation. AvSAM also shows the lowest spectral alignment (0.366), supporting the misalignment-based explanation of sharpness cancellation. InSAM reduces both average domain-wise sharpness (0.040→0.033) and the Cancellation Gap (0.020→0.012), but yields the highest sharpness dominance ratio (0.713), indicating that gap reduction can coincide with sharpness imbalance.

CGM suppresses imbalance and improves alignment. CGM achieves the lowest average domain-wise sharpness (0.030) while keeping average-loss sharpness low (0.019). Although its Cancellation Gap is comparable to that of InSAM (0.011 vs. 0.012), CGM reduces sharpness dominance (0.619 vs. 0.713) and improves spectral alignment (0.412). This indicates that compared with InSAM, CGM maintains a similarly small gap with less reliance on sharpness imbalance and stronger directional alignment. Figure 2 shows the training dynamics of these diagnostics, reflecting the results in Table 2: AvSAM maintains relatively large average domain-wise sharpness and Cancellation Gap, InSAM reduces the gap but exhibits the largest sharpness dominance, and CGM keeps the gap low while mitigating the dominance observed in InSAM.

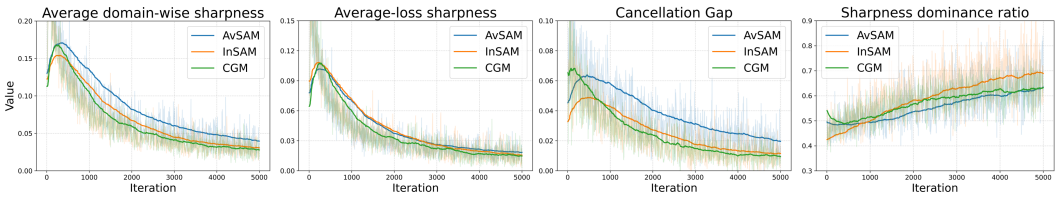


Figure 2: Diagnostic dynamics on PACS. For each leave-one-domain-out split, each method is trained once, and source-domain diagnostics are tracked during training; curves are averaged over splits. Faint curves show raw diagnostic values; thick curves show trends smoothed by an exponential moving average (smoothing factor 0.99).

Table 3: Domain-wise sharpness analysis on PACS with Sketch (S) as the target domain and Art (A), Cartoon (C), and Photo (P) as source domains. We report domain-wise sharpness estimates and maximum Hessian eigenvalues, averaged over five independent runs (mean ± standard error).

Metric	Method	A	C	P	Avg.
$\widehat{\mathcal{S}}_i^\rho$	AvSAM	$0.055_{\pm 0.013}$	$0.044_{\pm 0.016}$	$0.027_{\pm 0.010}$	0.042
	InSAM	$0.045_{\pm 0.012}$	$0.038_{\pm 0.010}$	$\mathbf{0.019}_{\pm 0.006}$	0.034
	CGM	$\mathbf{0.041}_{\pm 0.012}$	$\mathbf{0.032}_{\pm 0.007}$	$0.020_{\pm 0.003}$	**0.031**
$\lambda_{\max}(H_i)$	AvSAM	$22.3_{\pm 5.2}$	$18.3_{\pm 3.2}$	$10.3_{\pm 3.5}$	17.0
	InSAM	$13.9_{\pm 2.6}$	$9.7_{\pm 2.2}$	$2.9_{\pm 0.7}$	8.8
	CGM	$\mathbf{11.8}_{\pm 1.8}$	$\mathbf{8.0}_{\pm 1.4}$	$\mathbf{2.6}_{\pm 0.7}$	**7.5**

CGM effectively reduces domain-wise sharpness. While Table 2 shows that CGM achieves the lowest average domain-wise sharpness, we further analyze domain-wise sharpness on the PACS split with Sketch (S) as the target domain and Art (A), Cartoon (C), and Photo (P) as source domains. For each source domain, we report both the SAM-based sharpness estimate, $\widehat{\mathcal{S}}_i^\rho$, and the maximum Hessian eigenvalue, $\lambda_{\max}(H_i)$, which serves as a direct curvature-based sharpness measure. The maximum Hessian eigenvalue is approximated using power iteration with Hessian-vector products. Results are shown in Table 3. InSAM consistently improves over AvSAM across all source domains, and CGM further reduces domain-wise sharpness compared to InSAM. Although its $\widehat{\mathcal{S}}_i^\rho$ on Photo (P) is slightly higher than that of InSAM (0.020 vs. 0.019), CGM more effectively reduces sharpness on the relatively sharper Art (A) and Cartoon (C) domains. Moreover, CGM achieves the lowest $\lambda_{\max}(H_i)$ across all source domains. These results indicate that CGM reduces domain-wise sharpness at the individual-domain level.

Notes on interpretation. The averaged $\widehat{\mathcal{S}}_i^\rho$ values in Table 3 should not be used to recompute $\widehat{w}_{\max}$ in Table 2. Even when one domain dominates in each run, the identity of the sharpest domain may vary across runs. For example, sharpness patterns such as $(8, 1, 1)$ and $(1, 8, 1)$ both yield a large $\widehat{w}_{\max} = 0.8$, but their per-domain average becomes $(4.5, 4.5, 1)$, which appears less imbalanced. Therefore, recomputing $\widehat{w}_{\max}$ from averaged $\widehat{\mathcal{S}}_i^\rho$ can substantially underestimate sharpness dominance. This explains why Table 3 may appear less imbalanced than the $\widehat{w}_{\max}$ values reported in Table 2. In addition, $\widehat{\mathcal{S}}_i^\rho$ and $\lambda_{\max}(H_i)$ are not necessarily proportional: the former is a SAM-based sharpness estimate, whereas the latter is a local curvature measure. Finally, although the numerical gain in spectral alignment may appear moderate, such an improvement can be meaningful in a high-dimensional parameter space. The goal of InSAM and CGM is not to enforce perfect alignment, but to mitigate harmful misalignment enough to avoid flat-on-average yet sharp-per-domain solutions. Thus, the observed alignment gain is consistent with the intended role of CGM.

5.4 Ablation study

We analyze the sensitivity of CGM to its two method-specific hyperparameters: the imbalance penalty coefficient α and the warm-up ratio γ. Experiments are conducted on OfficeHome, varying one hyperparameter at a time while keeping all others fixed. Specifically, we use a learning rate of 10^{-5}, dropout rate of 0.5, weight decay of 10^{-4}, and perturbation radius $\rho = 0.05$. When varying α, we fix $\gamma = 0.1$; when varying γ, we fix $\alpha = 10^{-3}$.

Table 4: Computational cost on PACS using ResNet-50. We report median update time (s) and peak GPU memory (GiB) on an NVIDIA RTX A6000. D denotes the number of simulated domain-wise computational units, each consisting of a sub-batch of 32 examples.

	$D = 3$		$D = 9$		$D = 15$	
Method	Time	Memory	Time	Memory	Time	Memory
AvSAM	0.368	8.46	1.085	24.6	1.809	40.8
InSAM	0.383	3.54	1.149	4.70	1.922	5.86
CGM	0.386	3.54	1.160	4.70	1.940	5.86

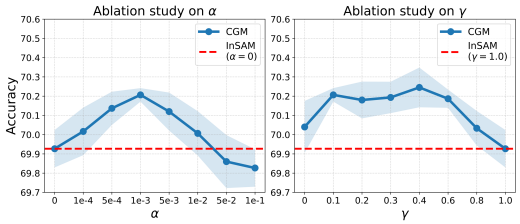


Figure 3: Ablation studies on OfficeHome for CGM hyperparameters α and γ. Thick lines and shaded regions indicate the mean and standard deviation over three independent runs, respectively.

Effect of α. The left panel of Figure 3 shows that moderate $\alpha > 0$ improves performance over InSAM ($\alpha = 0$), with the best result around $\alpha = 10^{-3}$. This indicates that controlling sharpness imbalance provides benefits beyond penalizing average domain-wise sharpness. However, overly large α degrades performance, suggesting that excessive imbalance regularization overconstrains optimization.

Effect of γ. The right panel of Figure 3 shows that using a warm-up phase consistently outperforms InSAM ($\gamma = 1.0$), with the best performance around $\gamma = 0.3$–0.6. This aligns with our analysis: early in training, domain-wise sharpness estimates are noisy, whereas after sufficient InSAM warm-up, optimization enters a regime where CGM is effective. When γ is too large, insufficient training remains for CGM to correct sharpness imbalance.

5.5 Computational cost analysis

For a fixed total batch size, AvSAM, InSAM, and CGM have comparable theoretical FLOPs, although the separate domain-wise passes of InSAM and CGM introduce execution overhead. To evaluate how this overhead scales, we simulate larger domain counts on PACS. For each of the three source domains, we split $32n$ examples into n sub-batches of size 32, yielding $D = 3n$ computational units rather than distinct domains. We test $D \in \{3, 9, 15\}$ with ResNet-50 on an RTX A6000, using the same total batch size across methods at each D. AvSAM processes the aggregated batch, whereas InSAM and CGM process sub-batches sequentially. As shown in Table 4, CGM requires only 1.05–$1.07\times$ the update time of AvSAM and adds less than 1% overhead over InSAM, excluding data loading and augmentation. At $D = 15$, peak GPU memory decreases from 40.8 GiB for AvSAM to 5.86 GiB for CGM through sequential processing.

6 Conclusion

Summary. We revisited SAM for domain generalization (DG) from a geometric perspective. We showed that applying SAM to the average loss (AvSAM) can suffer from *sharpness cancellation*: misaligned domain-wise dominant Hessian eigenvectors may make the average loss appear flat while individual domains remain sharp. We then introduced InSAM, which penalizes average domain-wise sharpness, and showed it decomposes into AvSAM plus a non-negative *Cancellation Gap*, which is minimized only under maximal alignment. We further identified a degeneracy where the Cancellation Gap becomes small due to domain-wise sharpness imbalance (a *bad shortcut*). To prevent this, we proposed CGM, which augments InSAM with a CV^2 regularizer to discourage sharpness imbalance. Across five DG benchmarks, CGM achieved the best average performance among strong baselines, with diagnostics confirming reduced average domain-wise sharpness and improved cross-domain alignment.

Limitations. A limitation of CGM is that its CV^2 regularizer treats all source domains symmetrically, encouraging balanced domain-wise sharpness. While natural without prior knowledge of domain difficulty or priority, this may be suboptimal when some domains should be emphasized more than others. Designing priority-aware variants that enable domain-dependent sharpness regularization is a promising direction for future work.

Acknowledgments and Disclosure of Funding

This work was supported by the Ministry of Science and ICT (MSIT), Republic of Korea, through the National Research Foundation of Korea (NRF) (2023R1A2C2005453, RS-2023-00218913) and the National IT Industry Promotion Agency (NIPA) (RS-2026-25621650). The authors declare no competing interests.

References

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

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

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

[4] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[5] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020.

[6] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2100–2110, 2019.

[7] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020.

[8] Yuge Shi, Jeffrey Seely, Philip HS Torr, Narayanaswamy Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

[9] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022.

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

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

[12] 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.

[13] 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.

[14] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124018, 2019.

[15] 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.

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

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

[18] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5189–5202, 2023.

[19] Seungjae Shin, HeeSun Bae, Byeonghu Na, Yoon-Yeong Kim, and Il-Chul Moon. Unknown domain inconsistency minimization for domain generalization. *arXiv preprint arXiv:2403.07329*, 2024.

[20] 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.

[21] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[22] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*, 2020.

[23] Xiran Wang, Jian Zhang, Lei Qi, and Yinghuan Shi. Balanced direction from multifarious choices: Arithmetic meta-learning for domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30577–30587, 2025.

[24] Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20558–20568, 2025.

[25] 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.

[26] 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*, pages 20247–20257, 2023.

[27] Binh M Le and Simon S Woo. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 188–199, 2024.

[28] 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*, pages 15349–15359, 2025.

[29] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *arXiv preprint arXiv:2106.02266*, 2021.

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

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

[32] 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*, pages 1657–1664, 2013.

[33] 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*, pages 5018–5027, 2017.

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

[35] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1406–1415. IEEE, 2019.

[36] 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.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[41] Marvin Mengxin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: A meta-learning approach for tackling group shift. 2020.

[42] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European conference on computer vision*, pages 124–140. Springer, 2020.

[43] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6502–6509, 2020.

[44] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[45] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9619–9628, 2021.

[46] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8690–8699, 2021.

[47] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34: 21189–21201, 2021.

Appendix for the Paper

"Beyond Average Flatness:
Domain-wise Flatness for Domain Generalization"

A Proofs and derivations

- A.1 Proof for Proposition 3.1
- A.2 Proof for Theorem 3.4
- A.3 Proof for Theorem 3.5
- A.4 Proof for Proposition 4.1
- A.5 Derivation of the CGM gradient

B Comparison with risk-based domain balancing

C Additional experimental details and results

- C.1 Experimental setups
- C.2 Coefficient dynamics and clipping effects
- C.3 Paired comparison with SAGM
- C.4 Compatibility with alternative sharpness measures
- C.5 Extended main results

A Proofs and derivations

Lemma A.1 (Quadratic maximizer)**.** *Let $H \succeq 0$ be symmetric, and suppose that its dominant eigenvalue $\lambda_{\max}(H) > 0$ is simple with unit dominant eigenvector v. Then*

$$\max_{\|\varepsilon\|_2 \le \rho} \frac{1}{2}\varepsilon^\top H \varepsilon = \frac{\rho^2}{2}\lambda_{\max}(H),$$

and the maximizers are exactly

$$\arg\max_{\|\varepsilon\|_2 \le \rho} \frac{1}{2}\varepsilon^\top H \varepsilon = \{\rho v, -\rho v\}.$$

Proof. For any $\varepsilon \neq 0$, write $\varepsilon = \|\varepsilon\|_2 u$ with $\|u\|_2 = 1$. By the variational characterization of the largest eigenvalue,

$$u^\top H u \le \lambda_{\max}(H).$$

Thus,

$$\frac{1}{2}\varepsilon^\top H \varepsilon = \frac{1}{2}\|\varepsilon\|_2^2 u^\top H u \le \frac{1}{2}\rho^2 \lambda_{\max}(H).$$

Equality is achieved by $\varepsilon = \rho v$ or $\varepsilon = -\rho v$. Since the dominant eigenvalue is simple, equality in the Rayleigh quotient occurs only for $u = \pm v$. Hence the maximizers are exactly $\{\rho v, -\rho v\}$. □

A.1 Proof for Proposition 3.1

Proof. Let

$$\mathcal{L}_{\mathrm{Av}}(\theta) := \frac{1}{D}\sum_{i \in \mathcal{I}} \mathcal{L}_i(\theta), \qquad H_{\mathrm{Av}}(\theta) := \nabla^2 \mathcal{L}_{\mathrm{Av}}(\theta) = \frac{1}{D}\sum_{i \in \mathcal{I}} H_i(\theta).$$

Since $\theta \in \Theta^\star$ is a local minimizer of every domain loss $\mathcal{L}_i$, we have $\nabla \mathcal{L}_i(\theta) = 0$ for all $i \in \mathcal{I}$. Therefore,

$$\nabla \mathcal{L}_{\mathrm{Av}}(\theta) = \frac{1}{D}\sum_{i \in \mathcal{I}} \nabla \mathcal{L}_i(\theta) = 0.$$

We first relate the average-loss sharpness to the curvature of the average loss. By the second-order Taylor expansion around θ, for $\|\varepsilon\|_2 \le \rho$,

$$\mathcal{L}_{\mathrm{Av}}(\theta + \varepsilon) - \mathcal{L}_{\mathrm{Av}}(\theta) = \frac{1}{2}\varepsilon^\top H_{\mathrm{Av}}(\theta)\varepsilon + o(\|\varepsilon\|_2^2).$$

Hence, as $\rho \to 0$,

$$\mathcal{S}^\rho_{\mathrm{Av}}(\theta) = \max_{\|\varepsilon\|_2 \le \rho} \left(\mathcal{L}_{\mathrm{Av}}(\theta + \varepsilon) - \mathcal{L}_{\mathrm{Av}}(\theta)\right) = \max_{\|\varepsilon\|_2 \le \rho} \frac{1}{2}\varepsilon^\top H_{\mathrm{Av}}(\theta)\varepsilon + o(\rho^2).$$

Since each $H_i(\theta)$ is positive semidefinite, their average $H_{\mathrm{Av}}(\theta)$ is also positive semidefinite. For any symmetric matrix A, the largest eigenvalue satisfies

$$\lambda_{\max}(A) = \max_{\|u\|_2 = 1} u^\top A u.$$

Therefore,

$$\max_{\|\varepsilon\|_2 \le \rho} \varepsilon^\top H_{\mathrm{Av}}(\theta)\varepsilon = \rho^2 \lambda_{\max}(H_{\mathrm{Av}}(\theta)),$$

and thus

$$\mathcal{S}^\rho_{\mathrm{Av}}(\theta) = \frac{\rho^2}{2}\lambda_{\max}(H_{\mathrm{Av}}(\theta)) + o(\rho^2).$$

This gives the approximation

$$\mathcal{S}^\rho_{\mathrm{Av}}(\theta) \approx \frac{\rho^2}{2}\lambda_{\max}(H_{\mathrm{Av}}(\theta))$$

for sufficiently small $\rho > 0$.

It remains to prove the spectral-alignment lower bound. For each domain, let

$$\lambda_i := \lambda_{\max}(H_i(\theta)).$$

Since $H_i(\theta)$ is symmetric and positive semidefinite, it admits an orthonormal eigenbasis. Let v_i be a unit dominant eigenvector with eigenvalue λ_i. Completing v_i to an orthonormal eigenbasis, we can write

$$H_i(\theta) = \lambda_i v_i v_i^\top + \sum_{k\geq 2} \lambda_{i,k} q_{i,k} q_{i,k}^\top,$$

where $\lambda_{i,k} \geq 0$ for all k because $H_i(\theta)$ is positive semidefinite. Therefore, for any unit vector u,

$$\begin{aligned} u^\top H_i(\theta) u &= \lambda_i (u^\top v_i)^2 + \sum_{k\geq 2} \lambda_{i,k} (u^\top q_{i,k})^2 \\ &\geq \lambda_i (u^\top v_i)^2 \\ &\geq \lambda_- (u^\top v_i)^2, \end{aligned}$$

where $\lambda_- := \min_{i\in\mathcal{I}} \lambda_i$.

Now define

$$A(\theta) := \frac{1}{D} \sum_{i\in\mathcal{I}} v_i v_i^\top.$$

By definition,

$$\mathcal{A}_{\text{spec}}(\theta) = \lambda_{\max}(A(\theta)).$$

Let $u_\star$ be a unit eigenvector of $A(\theta)$ corresponding to $\lambda_{\max}(A(\theta))$. Using the variational characterization of the largest eigenvalue, we have

$$\begin{aligned} \lambda_{\max}(H_{\text{Av}}(\theta)) &= \max_{\|u\|_2=1} u^\top H_{\text{Av}}(\theta) u \\ &\geq u_\star^\top H_{\text{Av}}(\theta) u_\star \\ &= \frac{1}{D} \sum_{i\in\mathcal{I}} u_\star^\top H_i(\theta) u_\star \\ &\geq \frac{1}{D} \sum_{i\in\mathcal{I}} \lambda_- (u_\star^\top v_i)^2 \\ &= \lambda_- u_\star^\top \left(\frac{1}{D} \sum_{i\in\mathcal{I}} v_i v_i^\top \right) u_\star \\ &= \lambda_- \, u_\star^\top A(\theta) u_\star \\ &= \lambda_- \, \lambda_{\max}(A(\theta)) \\ &= \lambda_- \, \mathcal{A}_{\text{spec}}(\theta). \end{aligned}$$

This proves the desired lower bound. □

A.2 Proof for Theorem 3.4

Proof. Let

$$B_\rho := \{\varepsilon : \|\varepsilon\|_2 \leq \rho\}, \qquad \Delta_i(\varepsilon) := \mathcal{L}_i(\theta+\varepsilon) - \mathcal{L}_i(\theta).$$

Then the domain-wise sharpness and the average-loss sharpness can be written as

$$\mathcal{S}_i^\rho(\theta) = \max_{\varepsilon\in B_\rho} \Delta_i(\varepsilon), \qquad \mathcal{S}_{\text{Av}}^\rho(\theta) = \max_{\varepsilon\in B_\rho} \frac{1}{D} \sum_{i\in\mathcal{I}} \Delta_i(\varepsilon).$$

We first prove non-negativity. For any fixed $\varepsilon \in B_\rho$ and any domain $i \in \mathcal{I}$, we have

$$\Delta_i(\varepsilon) \leq \max_{\varepsilon'\in B_\rho} \Delta_i(\varepsilon') = \mathcal{S}_i^\rho(\theta).$$

Averaging over domains gives

$$\frac{1}{D}\sum_{i\in\mathcal{I}}\Delta_i(\varepsilon) \le \frac{1}{D}\sum_{i\in\mathcal{I}}\mathcal{S}_i^\rho(\theta).$$

Since this inequality holds for every $\varepsilon \in B_\rho$, it also holds after maximizing the left-hand side over B_ρ:

$$\mathcal{S}_{\mathrm{Av}}^\rho(\theta) = \max_{\varepsilon\in B_\rho}\frac{1}{D}\sum_{i\in\mathcal{I}}\Delta_i(\varepsilon) \le \frac{1}{D}\sum_{i\in\mathcal{I}}\mathcal{S}_i^\rho(\theta).$$

Therefore,

$$\mathcal{C}_\rho(\theta) = \frac{1}{D}\sum_{i\in\mathcal{I}}\mathcal{S}_i^\rho(\theta) - \mathcal{S}_{\mathrm{Av}}^\rho(\theta) \ge 0.$$

Next, we characterize when the above inequality becomes equality. Since B_ρ is compact and the losses are continuous, the maximum defining $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$ is attained. Let

$$\varepsilon^\star \in \arg\max_{\varepsilon\in B_\rho}\frac{1}{D}\sum_{i\in\mathcal{I}}\Delta_i(\varepsilon).$$

Suppose that $\mathcal{C}_\rho(\theta) = 0$. Then

$$\begin{aligned}
0 &= \mathcal{C}_\rho(\theta)\\
&= \frac{1}{D}\sum_{i\in\mathcal{I}}\mathcal{S}_i^\rho(\theta) - \frac{1}{D}\sum_{i\in\mathcal{I}}\Delta_i(\varepsilon^\star)\\
&= \frac{1}{D}\sum_{i\in\mathcal{I}}\left(\mathcal{S}_i^\rho(\theta) - \Delta_i(\varepsilon^\star)\right).
\end{aligned}$$

Each term $\mathcal{S}_i^\rho(\theta) - \Delta_i(\varepsilon^\star)$ is non-negative by the definition of $\mathcal{S}_i^\rho(\theta)$. Since their average is zero, every term must be zero. Hence,

$$\Delta_i(\varepsilon^\star) = \mathcal{S}_i^\rho(\theta) \qquad \text{for all } i \in \mathcal{I}.$$

Thus, if $\mathcal{C}_\rho(\theta) = 0$, there exists a single perturbation $\varepsilon^\star$ that simultaneously maximizes all domain-wise loss increases.

Conversely, suppose there exists $\varepsilon^\star \in B_\rho$ such that

$$\varepsilon^\star \in \arg\max_{\varepsilon\in B_\rho}\Delta_i(\varepsilon) \qquad \text{for all } i \in \mathcal{I}.$$

Then

$$\frac{1}{D}\sum_{i\in\mathcal{I}}\mathcal{S}_i^\rho(\theta) = \frac{1}{D}\sum_{i\in\mathcal{I}}\Delta_i(\varepsilon^\star) \le \max_{\varepsilon\in B_\rho}\frac{1}{D}\sum_{i\in\mathcal{I}}\Delta_i(\varepsilon) = \mathcal{S}_{\mathrm{Av}}^\rho(\theta).$$

Together with the non-negativity result, this implies $\mathcal{C}_\rho(\theta) = 0$. Therefore,

$$\mathcal{C}_\rho(\theta) = 0$$

is equivalent to the existence of a common perturbation that maximizes all domain-wise loss increases simultaneously.

Now assume $\theta \in \Theta^\star$ and the local quadratic model in Assumption 3.3. Then, for each domain,

$$\Delta_i(\varepsilon) = \frac{1}{2}\varepsilon^\top H_i(\theta)\varepsilon.$$

Let $\lambda_i := \lambda_{\max}(H_i(\theta))$ and let v_i be the corresponding unit dominant eigenvector. Since $H_i(\theta) \succeq 0$ and the dominant eigenvalue is positive and simple, Lemma A.1 gives

$$\mathcal{S}_i^\rho(\theta) = \frac{\rho^2}{2}\lambda_i, \qquad \arg\max_{\varepsilon\in B_\rho}\Delta_i(\varepsilon) = \{\rho v_i, -\rho v_i\}.$$

We now prove the two directions of the equivalence.

First, suppose $\mathcal{C}_\rho(\theta) = 0$. From the equality condition above, there exists a perturbation $\varepsilon^\star$ that maximizes the quadratic loss increase of every domain. Therefore,

$$\varepsilon^\star \in \{\rho v_i,\, -\rho v_i\} \qquad \text{for all } i \in \mathcal{I}.$$

This implies that all domain-wise dominant eigenvectors are aligned up to sign:

$$v_1 = \pm v_2 = \cdots = \pm v_D.$$

Thus, there exists a unit vector v such that $v_i = \pm v$ for all $i \in \mathcal{I}$. Since $v_i v_i^\top = v v^\top$, we have

$$\frac{1}{D} \sum_{i \in \mathcal{I}} v_i v_i^\top = v v^\top.$$

The matrix $v v^\top$ has largest eigenvalue 1. Hence,

$$\mathcal{A}_{\text{spec}}(\theta) = \lambda_{\max}\left(\frac{1}{D} \sum_{i \in \mathcal{I}} v_i v_i^\top \right) = 1.$$

Therefore,

$$\mathcal{C}_\rho(\theta) = 0 \quad \Longrightarrow \quad \mathcal{A}_{\text{spec}}(\theta) = 1.$$

Conversely, suppose $\mathcal{A}_{\text{spec}}(\theta) = 1$. Define

$$A(\theta) := \frac{1}{D} \sum_{i \in \mathcal{I}} v_i v_i^\top.$$

For any unit vector u,

$$u^\top A(\theta) u = \frac{1}{D} \sum_{i \in \mathcal{I}} (u^\top v_i)^2 \le 1,$$

because both u and v_i are unit vectors. Since $\lambda_{\max}(A(\theta)) = \mathcal{A}_{\text{spec}}(\theta) = 1$, there exists a unit vector $u^\star$ such that

$$(u^\star)^\top A(\theta) u^\star = 1.$$

Thus,

$$\frac{1}{D} \sum_{i \in \mathcal{I}} \big((u^\star)^\top v_i \big)^2 = 1.$$

Each term in the average is at most 1, and the average is exactly 1. Therefore, every term must be equal to 1:

$$\big((u^\star)^\top v_i \big)^2 = 1 \qquad \text{for all } i \in \mathcal{I}.$$

Since $u^\star$ and v_i are unit vectors, this implies

$$v_i = \pm u^\star \qquad \text{for all } i \in \mathcal{I}.$$

Therefore, the perturbation

$$\varepsilon^\star := \rho u^\star$$

belongs to $\{\rho v_i,\, -\rho v_i\}$ for every domain i. Hence, $\varepsilon^\star$ maximizes every domain-wise quadratic loss increase simultaneously. By the equality condition proved above, we obtain

$$\mathcal{C}_\rho(\theta) = 0.$$

Therefore,

$$\mathcal{A}_{\text{spec}}(\theta) = 1 \quad \Longrightarrow \quad \mathcal{C}_\rho(\theta) = 0.$$

Combining the two directions gives

$$\mathcal{C}_\rho(\theta) = 0 \iff \mathcal{A}_{\text{spec}}(\theta) = 1.$$

This completes the proof. □

A.3 Proof for Theorem 3.5

Proof. Let

$$B_\rho := \{\varepsilon : \|\varepsilon\|_2 \le \rho\}, \qquad \Delta_i(\varepsilon) := \mathcal{L}_i(\theta+\varepsilon) - \mathcal{L}_i(\theta).$$

Then the domain-wise sharpness and the average-loss sharpness can be written as

$$\mathcal{S}_i^\rho(\theta) = \max_{\varepsilon \in B_\rho} \Delta_i(\varepsilon), \qquad \mathcal{S}_{\mathrm{Av}}^\rho(\theta) = \max_{\varepsilon \in B_\rho} \frac{1}{D}\sum_{i\in\mathcal{I}} \Delta_i(\varepsilon).$$

By the local quadratic model in Assumption 3.3, for each domain $i \in \mathcal{I}$,

$$\Delta_i(\varepsilon) = \frac{1}{2}\varepsilon^\top H_i(\theta)\varepsilon.$$

Since $H_i(\theta) \succeq 0$, we have

$$\Delta_i(\varepsilon) \ge 0 \qquad \text{for all } i \in \mathcal{I} \text{ and all } \varepsilon \in B_\rho.$$

Let

$$i^\star \in \arg\max_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta), \qquad \mathcal{S}_{\max}^\rho(\theta) = \mathcal{S}_{i^\star}^\rho(\theta).$$

Since B_ρ is compact and $\Delta_{i^\star}$ is continuous, there exists

$$\varepsilon^\star \in \arg\max_{\varepsilon\in B_\rho} \Delta_{i^\star}(\varepsilon).$$

Thus,

$$\Delta_{i^\star}(\varepsilon^\star) = \mathcal{S}_{i^\star}^\rho(\theta) = \mathcal{S}_{\max}^\rho(\theta).$$

Now evaluate the average-loss sharpness at this particular perturbation $\varepsilon^\star$. Since $\mathcal{S}_{\mathrm{Av}}^\rho(\theta)$ is the maximum of the average loss increase over B_ρ,

$$\begin{aligned}\mathcal{S}_{\mathrm{Av}}^\rho(\theta) &= \max_{\varepsilon\in B_\rho} \frac{1}{D}\sum_{i\in\mathcal{I}} \Delta_i(\varepsilon) \\ &\ge \frac{1}{D}\sum_{i\in\mathcal{I}} \Delta_i(\varepsilon^\star).\end{aligned}$$

Using the non-negativity of all domain-wise loss increases, we can keep only the contribution from domain $i^\star$:

$$\frac{1}{D}\sum_{i\in\mathcal{I}} \Delta_i(\varepsilon^\star) \ge \frac{1}{D}\Delta_{i^\star}(\varepsilon^\star) = \frac{1}{D}\mathcal{S}_{\max}^\rho(\theta).$$

Therefore,

$$\mathcal{S}_{\mathrm{Av}}^\rho(\theta) \ge \frac{1}{D}\mathcal{S}_{\max}^\rho(\theta).$$

Using the definition of the Cancellation Gap,

$$\mathcal{C}_\rho(\theta) = \frac{1}{D}\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) - \mathcal{S}_{\mathrm{Av}}^\rho(\theta),$$

we obtain

$$\begin{aligned}\mathcal{C}_\rho(\theta) &\le \frac{1}{D}\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) - \frac{1}{D}\mathcal{S}_{\max}^\rho(\theta) \\ &= \frac{1}{D}\left(\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta) - \mathcal{S}_{\max}^\rho(\theta)\right).\end{aligned}$$

By the definition of $w_{\max}(\theta)$,

$$w_{\max}(\theta) = \frac{\mathcal{S}_{\max}^\rho(\theta)}{\sum_{i\in\mathcal{I}} \mathcal{S}_i^\rho(\theta)}.$$

Equivalently,

$$\mathcal{S}^{\rho}_{\max}(\theta) = w_{\max}(\theta) \sum_{i \in \mathcal{I}} \mathcal{S}^{\rho}_i(\theta).$$

Substituting this into the previous inequality gives

$$\begin{aligned}\mathcal{C}_\rho(\theta) &\leq \frac{1}{D}\left(\sum_{i \in \mathcal{I}} \mathcal{S}^{\rho}_i(\theta) - w_{\max}(\theta) \sum_{i \in \mathcal{I}} \mathcal{S}^{\rho}_i(\theta)\right) \\ &= \frac{1}{D}\Big(\sum_{i \in \mathcal{I}} \mathcal{S}^{\rho}_i(\theta)\Big)\big(1 - w_{\max}(\theta)\big).\end{aligned}$$

This proves the desired upper bound. □

A.4 Proof for Proposition 4.1

Proof. For simplicity, write

$$S_i := \mathcal{S}^{\rho}_i(\theta), \qquad \mu := \mu(\theta) = \frac{1}{D}\sum_{i \in \mathcal{I}} S_i, \qquad \sigma^2 := \sigma^2(\theta) = \frac{1}{D}\sum_{i \in \mathcal{I}} (S_i - \mu)^2,$$

and let $\sigma := \sqrt{\sigma^2}$. Since $\varepsilon = 0$ is included in the perturbation ball, each sharpness value satisfies $S_i \geq 0$. By the assumption $\mu > 0$, the sharpness dominance ratio is well-defined:

$$w_{\max}(\theta) = \frac{S_{\max}}{\sum_{i \in \mathcal{I}} S_i} = \frac{S_{\max}}{D\mu}, \qquad S_{\max} := \max_{i \in \mathcal{I}} S_i.$$

If $D = 1$, then $w_{\max}(\theta) = 1$ and $\mathrm{CV}^2(\theta) = 0$, so the claimed inequality holds with equality. Hence, in what follows, assume $D \geq 2$.

Let $i^\star \in \arg\max_{i \in \mathcal{I}} S_i$, so that $S_{i^\star} = S_{\max}$. Define centered sharpness values

$$a_i := S_i - \mu.$$

By the definition of the mean,

$$\sum_{i \in \mathcal{I}} a_i = \sum_{i \in \mathcal{I}} (S_i - \mu) = \sum_{i \in \mathcal{I}} S_i - D\mu = 0.$$

Let

$$a_\star := a_{i^\star} = S_{\max} - \mu.$$

Since $S_{\max}$ is at least the average μ, we have $a_\star \geq 0$. Moreover,

$$\sum_{i \neq i^\star} a_i = -a_\star.$$

We now lower-bound the variance in terms of $a_\star$. Among the $D - 1$ centered values $\{a_i\}_{i \neq i^\star}$, their average is

$$\frac{1}{D-1}\sum_{i \neq i^\star} a_i = -\frac{a_\star}{D-1}.$$

Using the fact that a sum of squared deviations is always non-negative, we have

$$\sum_{i \neq i^\star} \left(a_i + \frac{a_\star}{D-1}\right)^2 \geq 0.$$

Expanding this inequality gives

$$\begin{aligned}0 &\leq \sum_{i \neq i^\star} a_i^2 + \frac{2a_\star}{D-1}\sum_{i \neq i^\star} a_i + (D-1)\frac{a_\star^2}{(D-1)^2} \\ &= \sum_{i \neq i^\star} a_i^2 - \frac{2a_\star^2}{D-1} + \frac{a_\star^2}{D-1} \\ &= \sum_{i \neq i^\star} a_i^2 - \frac{a_\star^2}{D-1}.\end{aligned}$$

Therefore,

$$\sum_{i \neq i^\star} a_i^2 \geq \frac{a_\star^2}{D-1}.$$

Using the definition of σ^2, we obtain

$$\begin{aligned} D\sigma^2 &= \sum_{i \in \mathcal{I}} a_i^2 \\ &= a_\star^2 + \sum_{i \neq i^\star} a_i^2 \\ &\geq a_\star^2 + \frac{a_\star^2}{D-1} \\ &= \frac{D}{D-1} a_\star^2. \end{aligned}$$

Dividing both sides by D yields

$$\sigma^2 \geq \frac{a_\star^2}{D-1}.$$

Since $a_\star \geq 0$, this implies

$$a_\star \leq \sqrt{D-1}\,\sigma.$$

Recalling that $a_\star = S_{\max} - \mu$, we get

$$S_{\max} \leq \mu + \sqrt{D-1}\,\sigma.$$

Finally, divide both sides by $D\mu > 0$:

$$\begin{aligned} w_{\max}(\theta) = \frac{S_{\max}}{D\mu} &\leq \frac{\mu + \sqrt{D-1}\,\sigma}{D\mu} \\ &= \frac{1 + \sqrt{D-1}\,\frac{\sigma}{\mu}}{D}. \end{aligned}$$

By the definition of the squared coefficient-of-variation,

$$\mathrm{CV}^2(\theta) = \frac{\sigma^2}{\mu^2},$$

and since $\mu > 0$, we have

$$\frac{\sigma}{\mu} = \sqrt{\mathrm{CV}^2(\theta)}.$$

Therefore,

$$w_{\max}(\theta) \leq \frac{1}{D}\left(1 + \sqrt{(D-1)\,\mathrm{CV}^2(\theta)}\right).$$

This proves the desired bound. □

A.5 Derivation of the CGM gradient

Proof. We derive Eq. (4). Recall the SAM-based sharpness estimate for domain i:

$$\widehat{\mathcal{S}}_i^\rho(\theta) := \mathcal{L}_i(\theta + \tilde{\varepsilon}_i) - \mathcal{L}_i(\theta), \qquad \tilde{\varepsilon}_i = \rho \frac{\nabla \mathcal{L}_i(\theta)}{\|\nabla \mathcal{L}_i(\theta)\|_2}.$$

Let

$$\widehat{\mu}(\theta) := \frac{1}{D} \sum_{i \in \mathcal{I}} \widehat{\mathcal{S}}_i^\rho(\theta), \qquad \widehat{\sigma}^2(\theta) := \frac{1}{D} \sum_{i \in \mathcal{I}} \left(\widehat{\mathcal{S}}_i^\rho(\theta) - \widehat{\mu}(\theta)\right)^2.$$

To avoid reducing CV^2 by increasing the mean sharpness, we stop gradients through the denominator $\widehat{\mu}^2(\theta)$. Denoting the stop-gradient operator by $\mathrm{sg}(\cdot)$, the practical surrogate is

$$\widehat{\mathcal{L}}_{\mathrm{CGM}}(\theta) := \mathcal{L}_{\mathrm{Av}}(\theta) + \frac{1}{D} \sum_{i \in \mathcal{I}} \widehat{\mathcal{S}}_i^\rho(\theta) + \alpha \frac{\widehat{\sigma}^2(\theta)}{\mathrm{sg}(\widehat{\mu}^2(\theta))}.$$

In the following derivation, we write

$$S_i := \widehat{\mathcal{S}}_i^{\rho}(\theta), \qquad \mu := \widehat{\mu}(\theta), \qquad \sigma^2 := \widehat{\sigma}^2(\theta)$$

for simplicity.

Under the standard SAM first-order approximation, we ignore gradients through the perturbation $\tilde{\varepsilon}_i$. Hence,

$$\nabla_\theta \mathcal{L}_i(\theta + \tilde{\varepsilon}_i) \approx \nabla \mathcal{L}_i(\theta + \tilde{\varepsilon}_i).$$

Therefore, the gradient of the sharpness estimate is

$$\begin{aligned}\nabla S_i &= \nabla_\theta \left(\mathcal{L}_i(\theta + \tilde{\varepsilon}_i) - \mathcal{L}_i(\theta)\right) \\ &\approx \nabla \mathcal{L}_i(\theta + \tilde{\varepsilon}_i) - \nabla \mathcal{L}_i(\theta).\end{aligned}$$

For brevity, define

$$g_i := \nabla \mathcal{L}_i(\theta), \qquad \tilde{g}_i := \nabla \mathcal{L}_i(\theta + \tilde{\varepsilon}_i).$$

Then

$$\nabla S_i \approx \tilde{g}_i - g_i.$$

We now compute the gradient of the variance term. Since

$$\mu = \frac{1}{D}\sum_{j \in \mathcal{I}} S_j, \qquad \sigma^2 = \frac{1}{D}\sum_{j \in \mathcal{I}} (S_j - \mu)^2,$$

we have

$$\frac{\partial \mu}{\partial S_i} = \frac{1}{D}.$$

Also,

$$\begin{aligned}\frac{\partial \sigma^2}{\partial S_i} &= \frac{\partial}{\partial S_i}\left[\frac{1}{D}\sum_{j \in \mathcal{I}} (S_j - \mu)^2\right] \\ &= \frac{2}{D}\sum_{j \in \mathcal{I}} (S_j - \mu)\left(\mathbf{1}\{j = i\} - \frac{\partial \mu}{\partial S_i}\right) \\ &= \frac{2}{D}\sum_{j \in \mathcal{I}} (S_j - \mu)\left(\mathbf{1}\{j = i\} - \frac{1}{D}\right).\end{aligned}$$

Since $\sum_{j \in \mathcal{I}} (S_j - \mu) = 0$, the second term vanishes, and thus

$$\frac{\partial \sigma^2}{\partial S_i} = \frac{2}{D}(S_i - \mu).$$

Because gradients through μ^2 in the denominator are stopped, we treat μ^2 as a constant during differentiation. Therefore,

$$\begin{aligned}\nabla\left(\frac{\sigma^2}{\mathrm{sg}(\mu^2)}\right) &= \sum_{i \in \mathcal{I}} \frac{1}{\mathrm{sg}(\mu^2)} \frac{\partial \sigma^2}{\partial S_i} \nabla S_i \\ &\approx \frac{1}{D}\sum_{i \in \mathcal{I}} \frac{2(S_i - \mu)}{\mu^2}(\tilde{g}_i - g_i).\end{aligned}$$

We can now assemble the full gradient. Since

$$\mathcal{L}_{\mathrm{Av}}(\theta) = \frac{1}{D}\sum_{i \in \mathcal{I}} \mathcal{L}_i(\theta),$$

we have

$$\nabla \mathcal{L}_{\mathrm{Av}}(\theta) = \frac{1}{D}\sum_{i \in \mathcal{I}} g_i.$$

Thus,

$$\nabla\widehat{\mathcal{L}}_{\mathrm{CGM}}(\theta) \approx \frac{1}{D}\sum_{i\in\mathcal{I}} g_i + \frac{1}{D}\sum_{i\in\mathcal{I}}(\tilde{g}_i - g_i) + \frac{\alpha}{D}\sum_{i\in\mathcal{I}}\frac{2(S_i-\mu)}{\mu^2}(\tilde{g}_i - g_i)$$
$$= \frac{1}{D}\sum_{i\in\mathcal{I}}\left[g_i + \left(1 + \frac{2\alpha}{\mu^2}(S_i-\mu)\right)(\tilde{g}_i - g_i)\right].$$

Define

$$c_i := 1 + \frac{2\alpha}{\mu^2}(S_i - \mu).$$

Then

$$g_i + c_i(\tilde{g}_i - g_i) = (1-c_i)g_i + c_i\tilde{g}_i.$$

Substituting back

$$S_i = \widehat{\mathcal{S}}_i^{\rho}(\theta), \qquad \mu = \widehat{\mu}(\theta),$$

we obtain

$$c_i = 1 + \frac{2\alpha}{\widehat{\mu}^2(\theta)}\left(\widehat{\mathcal{S}}_i^{\rho}(\theta) - \widehat{\mu}(\theta)\right).$$

Therefore,

$$\nabla\widehat{\mathcal{L}}_{\mathrm{CGM}}(\theta) \approx \frac{1}{D}\sum_{i\in\mathcal{I}}\left((1-c_i)\nabla\mathcal{L}_i(\theta) + c_i\nabla\mathcal{L}_i(\theta+\tilde{\varepsilon}_i)\right),$$

which is Eq. (4). □

B Comparison with risk-based domain balancing

Objective-level distinction. V-REx [11] and GroupDRO [21] address cross-domain robustness through domain/group risks. Let $\mathcal{L}_i(\theta)$ denote the empirical risk of domain i, and let $\mu_L = \frac{1}{D}\sum_{i=1}^{D}\mathcal{L}_i(\theta)$. Their objectives can be expressed as

$$\mathcal{L}_{\text{V-REx}}(\theta) = \frac{1}{D}\sum_{i=1}^{D}\mathcal{L}_i(\theta) + \frac{\beta}{D}\sum_{i=1}^{D}(\mathcal{L}_i(\theta) - \mu_L)^2,$$
$$\mathcal{L}_{\text{GroupDRO}}(\theta) = \max_{q\in\Delta_D}\sum_{i=1}^{D} q_i\mathcal{L}_i(\theta),$$

where Δ_D denotes the probability simplex over domains, treating each source domain as a group. V-REx encourages similar domain risks to promote robustness under distribution shifts, whereas GroupDRO directly targets worst-group risk.

In contrast, CGM regularizes domain-wise sharpness rather than domain risks. Let

$$S_i = \widehat{\mathcal{S}}_i(\theta) = \mathcal{L}_i(\theta + \tilde{\epsilon}_i) - \mathcal{L}_i(\theta), \qquad \mu_S = \frac{1}{D}\sum_{i=1}^{D} S_i,$$

where $\tilde{\epsilon}_i$ is the domain-wise SAM perturbation. CGM augments InSAM with a squared coefficient-of-variation penalty:

$$\mathcal{L}_{\text{CGM}}(\theta) = \frac{1}{D}\sum_{i=1}^{D}\mathcal{L}_i(\theta + \tilde{\epsilon}_i) + \alpha\frac{\frac{1}{D}\sum_{i=1}^{D}(S_i - \mu_S)^2}{\text{sg}(\mu_S^2)},$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation.

The sharpness-dispersion penalty serves a different purpose from risk balancing. It prevents a degenerate reduction of the Cancellation Gap driven by severe sharpness dominance. Thus, CGM does not impose identical sharpness across domains as a constraint; its primary objective remains reducing average domain-wise sharpness while avoiding this bad shortcut. Such a distinction is important because domains may exhibit different intrinsic difficulties, making strict sharpness equality undesirable.

Gradient-level distinction. The distinction also appears in the optimization updates. The gradient of the V-REx objective is

$$\nabla\mathcal{L}_{\text{V-REx}}(\theta) = \frac{1}{D}\sum_{i=1}^{D}\left[1 + 2\beta(\mathcal{L}_i(\theta) - \mu_L)\right]\nabla L_i(\theta).$$

Thus, V-REx modifies the coefficients of domain-risk gradients according to deviations in domain risks. Similarly, GroupDRO uses simplex-weighted domain-risk gradients, with greater emphasis on high-risk groups.

For CGM, using the stopped-gradient denominator, the gradient of the sharpness-dispersion penalty is

$$\nabla\left[\alpha\frac{\frac{1}{D}\sum_{i=1}^{D}(S_i - \mu_S)^2}{\text{sg}(\mu_S^2)}\right] = \frac{2\alpha}{D\mu_S^2}\sum_{i=1}^{D}(S_i - \mu_S)\nabla S_i.$$

Under the standard SAM approximation, which treats $\tilde{\epsilon}_i$ as fixed during differentiation,

$$\nabla S_i \approx \nabla\mathcal{L}_i(\theta + \tilde{\epsilon}_i) - \nabla\mathcal{L}_i(\theta).$$

Consequently, the CGM update can be written as

$$\nabla\mathcal{L}_{\text{CGM}}(\theta) \approx \frac{1}{D}\sum_{i=1}^{D}\Big[\nabla\mathcal{L}_i(\theta) + c_i\big(\nabla\mathcal{L}_i(\theta + \tilde{\epsilon}_i) - \nabla\mathcal{L}_i(\theta)\big)\Big],$$

where

$$c_i = 1 + \frac{2\alpha}{\mu_S^2}(S_i - \mu_S).$$

Unlike risk-based domain reweighting, CGM preserves uniform weighting of the base domain-risk gradients and modulates the within-domain SAM corrections according to relative sharpness. Domains with above-average sharpness receive larger coefficients on their SAM corrections, whereas those with below-average sharpness receive smaller coefficients. These coefficients are not simplex weights and need not lie between zero and one.

Interpretation. The two approaches also differ in their notions of worst-case optimization. GroupDRO targets the worst-performing group in terms of predictive risk, whereas SAM considers worst-case loss increases under local parameter perturbations. CGM does not perform worst-domain selection; instead, its smooth sharpness-dispersion penalty mitigates dominance by individual domains without introducing a max-domain objective.

Overall, V-REx and GroupDRO operate on domain-risk values, while CGM regulates the relative strength of domain-wise SAM corrections. More fundamentally, CGM's balancing mechanism is motivated by the geometry of the Cancellation Gap rather than risk equality or worst-group risk minimization.

C Additional experimental details and results

C.1 Experimental setups

Hyperparameter search space. For computational efficiency, recent DG baselines often use a discrete search space for non-method-specific hyperparameters, such as batch size, learning rate, weight decay, and dropout rate, instead of the original continuous search space used in DomainBed. Following SWAD [25] and subsequent baselines [17, 20, 23, 24, 36], we adopt the same discrete search space for non-method-specific hyperparameters. Among the baselines considered in our comparison, this protocol is also used by recent non-SAM-based methods such as GGA [24] and Arith [23], as well as most SAM-based baselines except GAM [26] and FAD [18]. Table 5 summarizes our hyperparameter search space.

Additional implementation details. Following recent practice, we train our implemented methods, i.e., AvSAM, InSAM, and CGM, for 15,000 iterations on DomainNet and 5,000 iterations on the other datasets. The reported results of the SAM-based baselines in our comparison also use the same iteration budget on DomainNet [17, 18, 20, 26, 36], as do recent non-SAM baselines such as GGA [24] and Arith [23]. Some earlier baselines report DomainNet results with 5,000 training iterations; therefore, comparisons on DomainNet across different iteration budgets should be interpreted with caution. Following DomainBed, we save checkpoints every 300 iterations for PACS, VLCS, OfficeHome, and TerraIncognita, and every 1,000 iterations for DomainNet. Depending on availability, we used either an RTX A6000 with 48GB memory or a TITAN V GPU with 12GB memory. Specifically, the AvSAM, InSAM, and CGM results in Table 1, as well as the reproduced SAGM results for the paired comparison, were obtained using TITAN V GPUs. For AvSAM and SAGM, we employed sequential domain-wise batch processing when out-of-memory errors occurred to reduce peak GPU memory usage. Gradients were accumulated across domains to preserve the original aggregated-loss update.

Table 5: Hyperparameter search space. $\mathrm{U}(a, b)$ denotes the uniform distribution over $[a, b]$.

Parameter	**DomainBed**	**Ours**
batch size	$2^{\mathrm{U}(3,5.5)}$	32
learning rate	$10^{\mathrm{U}(-5,-3.5)}$	$\{10^{-5}, 3\times10^{-5}, 5\times10^{-5}\}$
ResNet dropout	{0.0, 0.1, 0.5}	{0.0, 0.1, 0.5}
weight decay	$10^{\mathrm{U}(-6,-2)}$	$\{10^{-6}, 10^{-4}\}$
perturbation radius ρ	-	{0.01, 0.03, 0.05}
penalty coefficient α	-	$10^{\mathrm{U}(-3.5,-1.5)}$
warm-up ratio γ	-	{0.1, 0.2, 0.3}

C.2 Coefficient dynamics and clipping effects

Dynamics of clipped coefficients. We further visualize the dynamics of the clipped coefficient $\bar{c}_i$. Recall that $\bar{c}_i$ controls the interpolation between the standard gradient and the perturbed SAM gradient for domain i: larger $\bar{c}_i$ assigns more weight to the SAM gradient $\nabla\mathcal{L}_i(\theta+\tilde{\varepsilon}_i)$, while smaller $\bar{c}_i$ shifts the update toward the standard gradient $\nabla\mathcal{L}_i(\theta)$.

Figure 4 shows the clipped coefficient dynamics on the PACS split where Sketch (S) is used as the target domain and Art (A), Cartoon (C), and Photo (P) are used as source domains. We fix $\gamma = 0.1$, weight decay to 10^{-4}, dropout rate to 0.5, perturbation radius to $\rho = 0.03$, and learning rate to 3×10^{-5}, and vary only the penalty coefficient $\alpha \in \{0.0005, 0.001, 0.003, 0.01\}$. Each setting is trained once. Faint curves show raw clipped coefficient values, while thick curves show trends obtained by exponential moving average (EMA) smoothing with smoothing factor 0.99.

Because $\gamma = 0.1$ and training lasts for 5,000 iterations, the first 500 iterations correspond to the InSAM warm-up phase, where all coefficients are fixed to $\bar{c}_i = 1$. After the warm-up phase, larger α makes the coefficient more sensitive to centered sharpness differences. Since coefficients larger than 1 are clipped, this sensitivity mainly appears as larger downward deviations for flatter-than-average domains, resulting in lower EMA curves as α increases. For a fixed α, the coefficients are lower on relatively flatter domains, especially Photo, and higher on sharper domains such as Art and Cartoon.

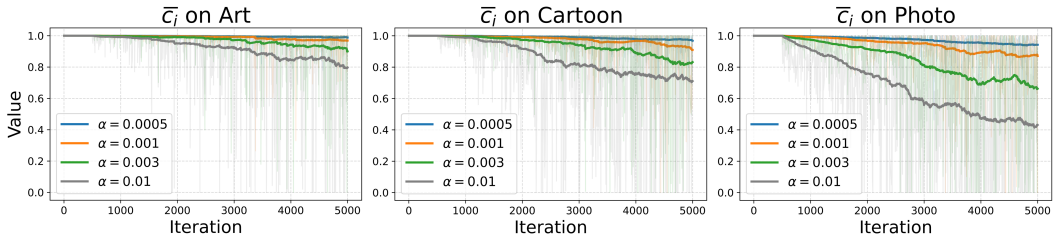


Figure 4: Dynamics of the clipped coefficients $\bar{c}_i$ on PACS with Sketch (S) as the target domain and Art (A), Cartoon (C), and Photo (P) as source domains. We vary the CGM penalty coefficient α while fixing $\gamma = 0.1$, weight decay 10^{-4}, dropout rate 0.5, $\rho = 0.03$, and learning rate 3×10^{-5}. Faint curves show raw coefficient values; thick curves show trends smoothed by an exponential moving average with smoothing factor 0.99.

Table 6: Coefficient clipping frequencies (%) and mean relative update difference on PACS. Statistics are computed over post-warm-up iterations. The frequency of unclipped coefficients is the complement of the upper and lower clipping frequencies.

	Upper clipping (%)			Lower clipping (%)			Mean
α	A	C	P	A	C	P	Δ_{clip}
5×10^{-4}	69.8	47.1	14.4	0.0	0.0	0.0	0.045
3×10^{-3}	70.3	48.8	14.5	0.2	1.2	3.5	0.184
10^{-2}	68.9	47.2	15.1	3.7	7.1	18.9	0.369

This is consistent with the CGM update rule: flatter domains receive reduced SAM-gradient weights, whereas sharper domains retain larger SAM-gradient weights.

Effect of coefficient clipping. CGM uses clipped coefficients $\bar{c}_i = \text{clip}(c_i, 0, 1)$ to keep each domain-wise update as a convex combination of the standard and perturbed SAM gradients. Without clipping, $c_i > 1$ extrapolates beyond the SAM gradient, whereas $c_i < 0$ reverses the SAM correction. We examine the practical effect of this heuristic on the PACS split with Art (A), Cartoon (C), and Photo (P) as source domains and Sketch (S) as the target domain. All statistics are computed over iterations after the warm-up phase, with $\gamma = 0.2$.

We measure the frequencies of upper ($c_i > 1$) and lower ($c_i < 0$) clipping. Since clipping frequency alone does not quantify the resulting change in the update, we also define the relative update difference as

$$\Delta_{\text{clip}} := \frac{\|g_{\text{clip}} - g_{\text{unclip}}\|_2}{\|g_{\text{unclip}}\|_2},$$

where g_{clip} and g_{unclip} denote the update directions computed using clipped and raw coefficients, respectively. As shown in Table 6, upper clipping occurs frequently across all tested values of α, reflecting which domains tend to have above-average sharpness. In contrast, lower clipping becomes more frequent as α increases, particularly for the relatively flatter Photo domain (from 0.0% to 18.9%). Despite frequent upper clipping, the mean relative update difference remains modest at $\alpha = 5 \times 10^{-4}$ (0.045), but increases to 0.184 and 0.369 as α increases. These results indicate that clipping frequency alone does not fully capture its effect on the update, which becomes more substantial as coefficient modulation strengthens.

C.3 Paired comparison with SAGM

To provide a controlled comparison with SAGM, which achieves the second-highest average accuracy in Table 1, we rerun SAGM and CGM under identical experimental settings using the same three random seeds. For each benchmark and seed, we average the target-domain accuracies over leave-one-domain-out splits and compute the paired difference between CGM and SAGM. The overall difference is computed by averaging the five benchmark-wise differences for each seed.

As shown in Table 7, the reproduced SAGM and CGM achieve average accuracies of 65.905% and 66.378%, respectively. CGM outperforms SAGM on average across all five benchmarks, with an overall paired improvement of 0.473 percentage points. The improvement is positive in 12

Table 7: Reproduced accuracies of SAGM and CGM (%) and paired accuracy differences (percentage points) under identical experimental settings and random seeds. The accuracy columns report means over three seeds, while paired differences are reported for each seed and as mean ± standard error. Positive differences indicate improvements by CGM.

Dataset	SAGM	CGM	Δ_0	Δ_1	Δ_2	Mean ± SE
PACS	86.837	87.476	+0.393	+1.432	+0.094	+0.640±0.405
VLCS	79.786	80.099	+0.763	-0.067	+0.243	+0.313±0.242
OfficeHome	69.868	70.461	+0.650	+0.451	+0.680	+0.593±0.072
TerraInc.	48.015	48.303	-0.107	-1.016	+1.986	+0.288±0.889
DomainNet	45.021	45.552	+0.361	+0.743	+0.491	+0.532±0.112
Overall	**65.905**	**66.378**	+0.412	+0.309	+0.699	**+0.473±0.117**

of the 15 benchmark-seed comparisons, although the differences vary across seeds, particularly on TerraIncognita. A two-sided paired t-test on the three seed-wise overall differences yields $p = 0.056$, with a 95% confidence interval of $[-0.029, 0.975]$ percentage points. Thus, while CGM shows positive mean improvements across all benchmarks, the limited number of seeds prevents a statistically significant conclusion at the 5% level.

C.4 Compatibility with alternative sharpness measures

We further examine whether CGM can be combined with alternative sharpness measures beyond the standard SAM sharpness. GAM [26] optimizes first-order sharpness, while FAD [18] combines zeroth- and first-order sharpness. We denote these sharpness measures as

$$\mathcal{R}_\rho^{(0)}(\theta) := \max_{\|\epsilon\|_2 \le \rho} \left(\mathcal{L}(\theta+\epsilon) - \mathcal{L}(\theta)\right),$$
$$\mathcal{R}_\rho^{(1)}(\theta) := \rho \max_{\|\epsilon\|_2 \le \rho} \|\nabla \mathcal{L}(\theta+\epsilon)\|_2,$$
$$\mathcal{R}_{\rho,\eta}^{\mathrm{FAD}}(\theta) := \eta \mathcal{R}_\rho^{(0)}(\theta) + (1-\eta)\mathcal{R}_\rho^{(1)}(\theta).$$

Here, $\mathcal{R}_\rho^{(0)}$ corresponds to the standard zeroth-order SAM sharpness, $\mathcal{R}_\rho^{(1)}$ is the first-order sharpness used in GAM, and $\mathcal{R}_{\rho,\eta}^{\mathrm{FAD}}$ is the combined sharpness measure used in FAD. The approximate gradients for these measures are provided in the original GAM and FAD papers; our goal here is to evaluate whether the AvSAM/InSAM/CGM framework is compatible with alternative sharpness measures. We follow the method-specific hyperparameter search space used in FAD [18]. Due to computational cost, we exclude DomainNet and report results on the remaining four benchmarks in Table 8.

Table 8: Compatibility of CGM with alternative sharpness measures. We combine AvSAM, InSAM, and CGM with the sharpness measures used in GAM [26] and FAD [18], and report target-domain accuracy on four DG benchmarks (mean ± standard error over 3 seeds).

Method	**PACS**	**VLCS**	**OfficeHome**	**TerraInc.**	**Avg.**
GAM + AvSAM	86.4$_{\pm 0.5}$	**78.9**$_{\pm 0.4}$	68.7$_{\pm 0.3}$	46.1$_{\pm 1.4}$	70.0
GAM + InSAM	86.5$_{\pm 0.6}$	78.8$_{\pm 0.4}$	68.9$_{\pm 0.2}$	**47.0**$_{\pm 1.0}$	70.3
GAM + CGM	**86.6**$_{\pm 0.6}$	78.8$_{\pm 0.4}$	**69.3**$_{\pm 0.2}$	46.7$_{\pm 0.9}$	**70.4**
FAD + AvSAM	87.2$_{\pm 0.4}$	79.1$_{\pm 0.4}$	69.5$_{\pm 0.2}$	46.9$_{\pm 1.3}$	70.7
FAD + InSAM	87.4$_{\pm 0.4}$	79.4$_{\pm 0.4}$	69.8$_{\pm 0.3}$	47.4$_{\pm 0.8}$	71.0
FAD + CGM	**87.5**$_{\pm 0.4}$	**79.5**$_{\pm 0.6}$	**70.0**$_{\pm 0.3}$	**47.8**$_{\pm 1.2}$	**71.2**

When combined with GAM, which focuses on first-order sharpness, InSAM and CGM do not substantially improve over AvSAM on average. This is consistent with the fact that our geometric analysis is developed for the standard zeroth-order sharpness induced by SAM, rather than for first-order sharpness. In contrast, when combined with FAD, which jointly optimizes zeroth- and first-order sharpness, InSAM and CGM both yield clear gains over AvSAM, and CGM further improves upon InSAM. This indicates that the proposed sharpness-imbalance regularization can be beneficial beyond standard SAM, especially when the sharpness measure includes a zeroth-order sharpness component.

C.5 Extended main results

Table 9 provides the complete set of results, including baselines omitted from the main text, whereas Tables 10–14 report per-domain accuracies for each dataset.

Table 9: We report the average accuracy over leave-one-domain-out splits (mean ± standard error over 3 seeds). Results marked with ∗ are obtained from our own experiments, while those marked with † and ‡ are taken from [17] and [18], respectively. Other baseline results are taken from the original papers.

Category	Method	PACS	VLCS	OfficeHome	TerraInc.	DomainNet	Avg.
Non-SAM DG baselines	MMD† [3]	84.7±0.5	77.5±0.9	66.3±0.1	42.2±1.6	23.4±9.5	58.8
	Mixstyle† [40]	85.2±0.3	77.9±0.5	60.4±0.3	44.0±0.7	34.0±0.1	60.3
	GroupDRO† [21]	84.4±0.8	76.7±0.6	66.0±0.7	43.2±1.1	33.3±0.2	60.7
	IRM† [10]	83.5±0.8	78.5±0.5	64.3±2.2	47.6±0.8	33.9±2.8	61.6
	ARM† [41]	85.1±0.4	77.6±0.3	64.8±0.3	45.5±0.3	35.5±0.2	61.7
	V-REx† [11]	84.9±0.6	78.3±0.2	66.4±0.6	46.4±0.6	33.6±2.9	61.9
	AND-mask† [22]	84.4±0.9	78.1±0.9	65.6±0.4	44.6±0.3	37.2±0.6	62.0
	DANN† [2]	83.6±0.4	78.6±0.4	65.9±0.6	46.7±0.5	38.3±0.1	62.6
	RSC† [42]	85.2±0.9	77.1±0.5	65.5±0.9	46.6±1.0	38.9±0.5	62.7
	Mixup† [43]	84.6±0.6	77.4±0.6	68.1±0.3	47.9±0.8	39.2±0.1	63.4
	MLDG† [4]	84.9±1.0	77.2±0.4	66.8±0.6	47.7±0.9	41.2±0.1	63.6
	ERM† [44]	85.5±0.2	77.3±0.4	66.5±0.3	46.1±1.8	43.8±0.1	63.9
	Fish† [8]	85.5±0.3	77.8±0.3	68.6±0.4	45.1±1.3	42.7±0.2	63.9
	Fishr [9]	85.5±0.4	77.8±0.1	67.8±0.1	47.4±1.6	41.7±0.0	64.0
	SelfReg [45]	85.6±0.4	77.8±0.9	67.9±0.7	47.0±0.3	41.5±0.0	64.0
	SagNet† [46]	86.3±0.2	77.8±0.5	68.1±0.1	48.6±1.0	40.3±0.1	64.2
	CORAL† [1]	86.2±0.3	78.8±0.6	68.7±0.3	47.7±1.0	41.5±0.1	64.5
	GGA [24]	86.4±0.5	78.7±0.8	67.0±0.3	48.5±1.1	44.4±0.2	65.0
	Arith [23]	86.5±0.3	79.4±0.8	69.4±0.3	48.1±1.2	41.5±0.2	65.0
	mDSDI† [47]	86.2±0.2	79.0±0.0	69.2±0.4	48.1±1.4	42.8±0.1	65.1
SAM-based DG baselines	GSAM† [36]	85.9±0.1	79.1±0.2	69.3±0.0	47.0±0.8	44.6±0.2	65.1
	GAM‡ [26]	86.1±0.6	78.5±0.4	68.2±1.0	45.2±0.6	43.8±0.1	64.4
	FAD‡ [18]	88.2±0.5	78.9±0.8	69.2±0.5	45.7±1.0	44.4±0.1	65.3
	SAGM† [17]	86.6±0.2	80.0±0.3	70.1±0.4	48.8±0.9	45.0±0.2	66.1
	DISAM [20]	87.1±0.2	79.9±0.5	70.3±0.2	46.6±0.8	45.4±0.2	65.9
	AvSAM∗	86.5±0.4	78.9±0.3	69.5±0.1	45.9±0.9	44.8±0.1	65.1
	InSAM∗	87.1±0.3	79.7±0.2	70.0±0.2	47.1±0.9	45.2±0.1	65.8
	CGM∗	87.5±0.5	80.1±0.1	70.5±0.1	48.3±0.9	45.6±0.0	66.4

Table 10: Out-of-domain accuracies (%) on PACS.

Category	Method	A	C	P	S	Avg.
Non-SAM DG baselines	MMD†	86.1±1.4	79.4±0.9	96.6±0.2	76.5±0.5	84.7
	Mixstyle†	86.8±0.5	79.0±1.4	96.6±0.1	78.5±2.3	85.2
	GroupDRO†	83.5±0.9	79.1±0.6	96.7±0.3	78.3±2.0	84.4
	IRM†	84.8±1.3	76.4±1.1	96.7±0.6	76.1±1.0	83.5
	ARM†	86.8±0.6	76.8±0.5	97.4±0.3	79.3±1.2	85.1
	V-REx†	86.0±1.6	79.1±0.6	96.9±0.5	77.7±1.7	84.9
	DANN†	86.4±0.8	77.4±0.8	97.3±0.4	73.5±2.3	83.6
	RSC†	85.4±0.8	79.7±1.8	97.6±0.3	78.2±1.2	85.2
	MLDG†	85.5±1.4	80.1±1.7	97.4±0.3	76.6±1.1	84.9
	ERM†	84.7±0.4	80.8±0.6	97.2±0.3	79.3±1.0	85.5
	Fishr	88.4±0.2	78.7±0.7	97.0±0.1	77.8±2.0	85.5
	SelfReg	87.9±1.0	79.4±1.4	96.8±0.7	78.3±1.2	85.6
	SagNet†	87.4±1.0	80.7±0.6	97.1±0.1	80.0±0.4	86.3
	CORAL†	88.3±0.2	80.0±0.5	97.5±0.3	78.8±1.3	86.2
	GGA	86.5±1.8	81.2±3.0	97.1±0.9	80.8±0.9	86.4
	Arith	85.9±0.5	81.3±1.0	97.1±0.5	81.8±1.0	86.5
	mDSDI†	87.7±0.4	80.4±0.7	98.1±0.3	78.4±1.2	86.2
SAM-based DG baselines	GSAM†	86.9±0.1	80.4±0.2	97.5±0.0	78.7±0.8	85.9
	GAM‡	85.9±0.9	81.3±1.6	98.2±0.4	79.0±2.1	86.1
	FAD‡	88.5±0.5	83.0±0.8	98.4±0.2	82.8±0.9	88.2
	SAGM†	87.4±0.2	80.2±0.3	98.0±0.2	80.8±0.6	86.6
	DISAM	87.1±0.4	81.9±0.5	96.2±0.3	83.1±0.7	87.1
	AvSAM*	86.8±0.2	81.1±0.9	97.0±0.7	81.2±1.0	86.5
	InSAM*	88.2±0.8	81.4±1.0	96.6±0.5	82.3±1.0	87.1
	CGM*	88.3±0.3	82.1±1.2	97.2±0.3	82.2±0.7	87.5

Table 11: Out-of-domain accuracies (%) on VLCS.

Category	Method	C	L	S	V	Avg.
Non-SAM DG baselines	MMD†	97.7±0.1	64.0±1.1	72.8±0.2	75.3±3.3	77.5
	Mixstyle†	98.6±0.3	64.5±1.1	72.6±0.5	75.7±1.7	77.9
	GroupDRO†	97.3±0.3	63.4±0.9	69.5±0.8	76.7±0.7	76.7
	IRM†	98.6±0.1	64.9±0.9	73.4±0.6	77.3±0.9	78.5
	ARM†	98.7±0.2	63.6±0.7	71.3±1.2	76.7±0.6	77.6
	V-REx†	98.4±0.3	64.4±1.4	74.1±0.4	76.2±1.3	78.3
	DANN†	99.0±0.3	65.1±1.4	73.1±0.3	77.2±0.6	78.6
	RSC†	97.9±0.1	62.5±0.7	72.3±1.2	75.6±0.8	77.1
	MLDG†	97.4±0.2	65.2±0.7	71.0±1.4	75.3±1.0	77.2
	ERM†	98.0±0.3	64.7±1.2	71.4±1.2	75.2±1.6	77.3
	Fishr	98.9±0.3	64.0±0.5	71.5±0.2	76.8±0.7	77.8
	SelfReg	96.7±0.4	65.2±1.2	73.1±1.3	76.2±0.7	77.8
	SagNet†	97.9±0.4	64.5±0.5	71.4±1.3	77.5±0.5	77.8
	CORAL†	98.3±0.1	66.1±1.2	73.4±0.3	77.5±1.2	78.8
	GGA	98.4±0.2	65.4±0.1	73.8±1.6	77.4±1.9	78.7
	Arith	98.7±0.3	64.3±0.8	76.0±0.9	78.6±1.0	79.4
	mDSDI†	97.6±0.1	66.4±0.4	74.0±0.6	77.8±0.7	79.0
SAM-based DG baselines	GSAM†	98.7±0.3	64.9±0.2	74.3±0.0	78.5±0.8	79.1
	GAM‡	98.8±0.6	65.1±1.2	72.9±1.0	77.2±1.9	78.5
	FAD‡	99.1±0.5	66.8±0.9	73.6±1.0	76.1±1.3	78.9
	SAGM†	99.0±0.2	65.2±0.4	75.1±0.3	80.7±0.8	80.0
	DISAM	99.3±0.0	66.3±0.5	81.0±0.1	73.2±0.1	79.9
	AvSAM*	99.0±0.3	64.7±0.3	72.5±0.9	79.5±0.4	78.9
	InSAM*	99.2±0.0	65.9±0.3	73.4±0.2	80.2±0.1	79.7
	CGM*	99.3±0.0	66.0±0.2	73.6±0.2	81.6±0.2	80.1

Table 12: Out-of-domain accuracies (%) on OfficeHome.

Category	Method	A	C	P	R	Avg.
Non-SAM DG baselines	MMD†	60.4±0.2	53.3±0.3	74.3±0.1	77.4±0.6	66.3
	Mixstyle†	51.1±0.3	53.2±0.4	68.2±0.7	69.2±0.6	60.4
	GroupDRO†	60.4±0.7	52.7±1.0	75.0±0.7	76.0±0.7	66.0
	IRM†	58.9±2.3	52.2±1.6	72.1±2.9	74.0±2.5	64.3
	ARM†	58.9±0.8	51.0±0.5	74.1±0.1	75.2±0.3	64.8
	V-REx†	60.7±0.9	53.0±0.9	75.3±0.1	76.6±0.5	66.4
	DANN†	59.9±1.3	53.0±0.3	73.6±0.7	76.9±0.5	65.9
	RSC†	60.7±1.4	51.4±0.3	74.8±1.1	75.1±1.3	65.5
	MLDG†	61.5±0.9	53.2±0.6	75.0±1.2	77.5±0.4	66.8
	ERM†	61.3±0.7	52.4±0.3	75.8±0.1	76.6±0.3	66.5
	Fishr	62.4±0.5	54.4±0.4	76.2±0.5	78.3±0.1	67.8
	SelfReg	63.6±1.4	53.1±1.0	76.9±0.4	78.1±0.4	67.9
	SagNet†	63.4±0.2	54.8±0.4	75.8±0.4	78.3±0.3	68.1
	CORAL†	65.3±0.4	54.4±0.5	76.5±0.1	78.4±0.5	68.7
	GGA	61.7±0.1	52.5±0.5	77.1±1.3	77.0±0.1	67.0
	Arith	64.6±0.8	55.3±0.9	78.3±0.4	79.4±0.6	69.4
	mDSDI†	68.1±0.3	52.1±0.4	76.0±0.2	80.4±0.2	69.2
SAM-based DG baselines	GSAM†	64.9±0.1	55.2±0.2	77.8±0.0	79.2±0.0	69.3
	GAM‡	63.0±1.2	49.8±0.5	77.6±0.6	82.4±1.0	68.2
	FAD‡	63.5±1.0	50.3±0.8	78.0±0.4	85.0±0.6	69.2
	SAGM†	65.4±0.4	57.0±0.3	78.0±0.3	80.0±0.2	70.1
	DISAM	65.8±0.2	55.6±0.2	79.2±0.2	80.6±0.1	70.3
	AvSAM*	64.3±0.1	57.1±0.1	77.4±0.4	78.9±0.1	69.5
	InSAM*	64.6±0.2	57.9±0.4	78.1±0.4	79.6±0.2	70.0
	CGM*	65.1±0.1	58.7±0.1	78.2±0.1	79.9±0.2	70.5

Table 13: Out-of-domain accuracies (%) on TerraIncognita.

Category	Method	L100	L38	L43	L46	Avg.
Non-SAM DG baselines	MMD†	41.9±3.0	34.8±1.0	57.0±1.9	35.2±1.8	42.2
	Mixstyle†	54.3±1.1	34.1±1.1	55.9±1.1	31.7±2.1	44.0
	GroupDRO†	41.2±0.7	38.6±2.1	56.7±0.9	36.4±2.1	43.2
	IRM†	54.6±1.3	39.8±1.9	56.2±1.8	39.6±0.8	47.6
	ARM†	49.3±0.7	38.3±2.4	55.8±0.8	38.7±1.3	45.5
	V-REx†	48.2±4.3	41.7±1.3	56.8±0.8	38.7±3.1	46.4
	DANN†	51.1±3.5	40.6±0.6	57.4±0.5	37.7±1.8	46.7
	RSC†	50.2±2.2	39.2±1.4	56.3±1.4	40.8±0.6	46.6
	MLDG†	54.2±3.0	44.3±1.1	55.6±0.3	36.9±2.2	47.7
	ERM†	49.8±4.4	42.1±1.4	56.9±1.8	35.7±3.9	46.1
	Fishr	50.2±3.9	43.9±0.8	55.7±2.2	39.8±1.9	47.4
	SelfReg	48.8±0.9	41.3±1.8	57.3±0.7	40.6±0.9	47.0
	SagNet†	53.0±2.9	43.0±2.5	57.9±0.6	40.4±1.3	48.6
	CORAL†	51.6±2.4	42.2±1.0	57.0±1.0	39.8±2.9	47.7
	GGA	50.9±2.2	42.5±1.0	59.7±1.4	41.5±3.5	48.5
	Arith	52.3±1.9	42.4±2.5	57.5±1.3	40.2±2.3	48.1
	mDSDI†	53.2±3.0	43.3±1.0	56.7±0.5	39.2±1.3	48.1
SAM-based DG baselines	GSAM†	50.8±0.1	39.3±0.2	59.6±0.0	38.2±0.8	47.0
	GAM‡	42.2±2.6	42.9±1.7	60.2±1.8	35.5±0.7	45.2
	FAD‡	44.3±2.2	43.5±1.7	60.9±2.0	34.1±0.5	45.7
	SAGM†	54.8±1.3	41.4±0.8	57.7±0.6	41.3±0.4	48.8
	DISAM	46.2±2.9	41.6±0.1	58.0±0.5	40.5±2.2	46.6
	AvSAM*	49.4±2.1	39.7±0.9	56.7±0.5	37.7±0.5	45.9
	InSAM*	50.8±2.0	41.5±1.1	57.7±0.6	38.5±0.9	47.1
	CGM*	54.3±1.4	43.2±1.9	57.4±0.5	38.2±0.9	48.3

Table 14: Out-of-domain accuracies (%) on DomainNet.

Category	Method	clip	info	paint	quick	real	sketch	Avg.
Non-SAM DG baselines	MMD†	32.1±13.3	11.0±4.6	26.8±11.3	8.7±2.1	32.7±13.8	28.9±11.9	23.4
	Mixstyle†	51.9±0.4	13.3±0.2	37.0±0.5	12.3±0.1	46.1±0.3	43.4±0.4	34.0
	GroupDRO†	47.2±0.5	17.5±0.4	33.8±0.5	9.3±0.3	51.6±0.4	40.1±0.6	33.3
	IRM†	48.5±2.8	15.0±1.5	38.3±4.3	10.9±0.5	48.2±5.2	42.3±3.1	33.9
	ARM†	49.7±0.3	16.3±0.5	40.9±1.1	9.4±0.1	53.4±0.4	43.5±0.4	35.5
	V-REx†	47.3±3.5	16.0±1.5	35.8±4.6	10.9±0.3	49.6±4.9	42.0±3.0	33.6
	DANN†	53.1±0.2	18.3±0.1	44.2±0.7	11.8±0.1	55.5±0.4	46.8±0.6	38.3
	RSC†	55.0±1.2	18.3±0.5	44.4±0.6	12.2±0.2	55.7±0.7	47.8±0.9	38.9
	MLDG†	59.1±0.2	19.1±0.3	45.8±0.7	13.4±0.3	59.6±0.2	50.2±0.4	41.2
	ERM†	62.8±0.4	20.2±0.3	50.3±0.3	13.7±0.5	63.7±0.2	52.1±0.5	43.8
	Fishr	58.2±0.5	20.2±0.2	47.7±0.3	12.7±0.2	60.3±0.2	50.8±0.1	41.7
	SelfReg	58.5±0.1	20.7±0.1	47.3±0.3	13.1±0.3	58.2±0.2	51.1±0.3	41.5
	SagNet†	57.7±0.3	19.0±0.2	45.3±0.3	12.7±0.5	58.1±0.5	48.8±0.2	40.3
	CORAL†	59.2±0.1	19.7±0.2	46.6±0.3	13.4±0.4	59.8±0.2	50.1±0.6	41.5
	GGA	63.7±0.2	21.3±0.3	50.4±0.1	14.1±0.2	63.8±0.4	53.5±0.3	44.4
	Arith	59.0±0.4	19.7±0.2	47.0±0.3	12.7±0.3	59.4±0.4	51.1±0.7	41.5
	mDSDI†	62.1±0.3	19.1±0.4	49.4±0.4	12.8±0.7	62.9±0.3	50.4±0.4	42.8
SAM-based DG baselines	GSAM†	64.2±0.3	20.8±0.2	50.9±0.0	14.4±0.8	63.5±0.2	53.9±0.2	44.6
	GAM‡	63.0±0.5	20.2±0.2	50.3±0.1	13.2±0.3	64.5±0.2	51.6±0.5	43.8
	FAD‡	64.1±0.3	21.9±0.2	50.6±0.3	14.2±0.4	63.6±0.1	52.2±0.2	44.4
	SAGM†	64.9±0.2	21.1±0.3	51.5±0.2	14.8±0.2	64.1±0.2	53.6±0.2	45.0
	DISAM	65.9±0.2	20.7±0.2	51.7±0.3	16.6±0.3	62.8±0.5	54.8±0.4	45.4
	AvSAM*	64.8±0.1	20.9±0.1	50.4±0.1	15.4±0.2	63.1±0.1	54.5±0.1	44.8
	InSAM*	64.6±0.1	21.5±0.2	51.3±0.2	15.7±0.1	63.7±0.3	54.5±0.3	45.2
	CGM*	65.1±0.1	21.3±0.1	52.1±0.1	15.8±0.1	64.1±0.1	54.9±0.1	45.6

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of this paper are discussed in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The complete proofs and derivations are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the code at `https://github.com/SeungjunChoi1/CGM`, including scripts to reproduce the main experimental results for our method. All datasets used are publicly available benchmarks.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed experimental settings in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In Section 5, we report the standard error over three runs for the main results and over five runs for the diagnostic analysis; for the ablation study, we visualize the standard deviation over three runs. For our experiments, standard errors are computed as the sample standard deviation across independent runs divided by the square root of the number of runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the relevant details about the computer resources and computational costs in Section 5.5 and Appendix C.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and confirm that the research presented in this paper complies with it in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The goal of this paper is to improve domain generalization for classification tasks. The work is primarily foundational, and we do not anticipate significant societal impact; in particular, we do not foresee any negative impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We used widely adopted benchmark datasets for DG, which doesn't have any risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used widely adopted benchmark datasets for domain generalization and cited the corresponding papers for each dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only to assist with writing, editing, and formatting. They did not contribute to the core methodology or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.