A Flat Minima Perspective on Understanding Augmentations and Model Robustness

Weebum Yoo[1], **Sung Whan Yoon**[1,2*]

[1]Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea
[2]Department of Electrical Engineering, Ulsan National Institute of Science and Technology (UNIST), Republic of Korea
{weebum, shyoon8}@unist.ac.kr

Abstract

Model robustness indicates a model's capability to generalize well on unforeseen distributional shifts, including data corruptions and adversarial attacks. Data augmentation is one of the most prevalent and effective ways to enhance robustness. Despite the great success of the diverse augmentations in different fields, a unified theoretical understanding of their efficacy in improving model robustness is lacking. We theoretically reveal a general condition for label-preserving augmentations to bring robustness to diverse distribution shifts through the lens of flat minima and generalization bound, which de facto turns out to be strongly correlated with robustness against different distribution shifts in practice. Unlike most earlier works, our theoretical framework accommodates all the label-preserving augmentations and is not limited to particular distribution shifts. We substantiate our theories through different simulations on the existing common corruption and adversarial robustness benchmarks based on the CIFAR and ImageNet datasets.

Code — https://github.com/pyoo96/aug-flatmin-robustness

Introduction

Model robustness, which is a critical factor of deep models in applications requiring high reliability, such as autonomous vehicles and medical diagnosis, entails maintaining performance against data distribution shifts. In the past decade, data augmentation has been widely used as a popular and pragmatic technique to enhance the model performance, as well as the robustness against data corruption, adversarial attacks, or even domain shifts (Xu et al. 2023; Zhao et al. 2020; Haddouche et al. 2025; Hendrycks et al. 2021a). The intuition of its efficacy relies on the belief that augmentations enrich the training data distribution, which allows models to easily extrapolate to unseen data distributions, which is the so-called generalization capability.

Despite the utility of augmentations, a unified theory formally clarifying how augmentations can enhance model robustness against diverse distribution shifts has not been well established. Prior analyses are mostly confined to either particular augmentations or adversarial robustness (Zhao et al. 2020; Rebuffi et al. 2021b; Najafi et al. 2019; Rebuffi et al. 2021a). Although some studies step forward from the aforementioned works, their analyses rely solely on empirical observations without a theoretical rationale, or the focus is exclusively on certain categories of distribution shifts (Zhang et al. 2024; Gao et al. 2020; Li et al. 2021).

In this paper, we offer a series of theoretical insights, including a sufficient condition (termed the PSA condition), that explains how label-preserving data augmentations can bolster model robustness against general distributional shifts through the lens of flat minima. Our analysis has two main branches: **i)** First, we mathematically bridge the general form of label-preserving augmentations to the improved generalization bound. To give a brief sketch of our development, we start to demonstrate the equivalence between the input space region covered by the augmented samples and the corresponding parameter space region with the same loss values (formalized by Theorem 1). Based on the equivalence, we then claim that the augmentation satisfying the PSA condition flattens the loss surface on the parameter space (Theorem 2), finally reaching to the improved generalization bound against distribution shifts via leveraging the flattened loss surface (Theorem 3). **ii)** Next, we validate our theoretical findings by evaluating existing augmentation methods across different robustness benchmarks, encompassing data distribution shifts caused by common corruptions and adversarial attacks. Our findings show that when augmentations have non-negligible sample coverage near the original image–which aligns with the PSA condition–they consistently enhance model robustness. In contrast, when augmentations fail to improve robustness, they exhibit negligible density near the original image, indicating that PSA serves as both a sufficient condition for robustness and a factor highly correlated with robustness.

Preliminaries

Basic Notations

Let us consider an input $x \in \mathbb{R}^n$ from input space $\mathcal{X}$, which is paired with a target label $y \in \mathbb{R}^c$ from label space $\mathcal{Y}$, where n and c are the dimensions of the input space and the label space. A model $f(\cdot;\theta) : \mathbb{R}^n \to \mathbb{R}^c$ parameterized by $\theta \in \mathbb{R}^p$ maps a given input to the estimated label, where p is the dimension of the model parameter space. The loss

*Corresponding Author

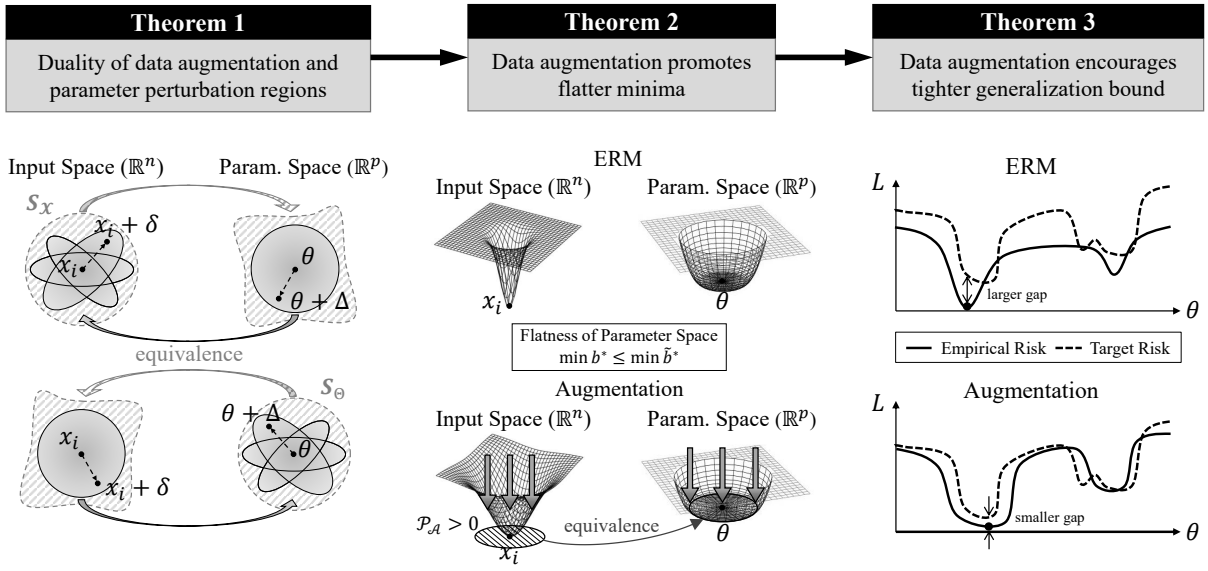

Figure 1: A conceptual overview of how augmentations can relate to model robustness. Theorem 1 shows that the neighboring region of an input datapoint can be mapped to an equivalent region in the parameter space (and vice versa). Theorem 2 demonstrates that ensuring a nonzero probability for an augmentation within this neighborhood (PSA condition) promotes flatter minima in parameter space by effectively regulating the proximal regions of both the input and parameter spaces. Finally, Theorem 3 presents a generalization bound for models trained with augmentations, insisting that augmentations with dense representations near the original image foster a tighter generalization gap.

function $\mathcal{L}(\cdot, \cdot)$ quantifies how far the predicted label $f(x; \theta)$ deviates from the true label y, expressed as $\mathcal{L}\big(f(x; \theta), y\big)$. Given a dataset $\{(x_i, y_i)\}_{i=1}^N$ of N samples drawn from data distribution $\mathcal{D}$, the true risk and the empirical risk are:

$$\mathcal{E}_\mathcal{D}(\theta) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}(f(x;\theta), y)\big] \tag{1}$$

$$\hat{\mathcal{E}}_\mathcal{D}(\theta) := \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f(x_i;\theta), y_i) \tag{2}$$

respectively, where $\mathcal{E}_\mathcal{D}(\theta)$ and $\hat{\mathcal{E}}_\mathcal{D}(\theta)$ denote the true and the empirical risk.

Data Augmentation Data augmentation $\mathcal{A}(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ augments a given input x to augmented input $\tilde{x} := \mathcal{A}(x)$. Let us further represent it with the difference between the original input and the augmented one: $\tilde{x} := \mathcal{A}(x) = x + \delta$, where $\delta \in \mathbb{R}^n$. [1] Also, we define $\mathcal{P}_\mathcal{A}(\tilde{x}|x)$ as the probability density function of an augmented sample $\tilde{x}$ given x. Finally, we let $\tilde{\mathcal{D}} := \{(\mathcal{A}(x), y) : (x, y) \sim \mathcal{D}\}$ represent the augmented dataset.

Model Flatness

Definition *Model flatness* characterizes the extent of change in the model's loss values across proximate points in the parameter space. When the loss rapidly changes around the found minima, it indicates that the model is located at *sharp minima*. Otherwise, it denotes *flat minima* when the loss varies smoothly. The change of losses around the model parameters can be formalized as follows:

$$\max_{\|\Delta\|\leq\gamma} \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}(f(x;\theta+\Delta), y) - \mathcal{L}(f(x;\theta), y)\big], \tag{3}$$

where $\Delta \in \mathbb{R}^p$ is the perturbation around model parameter θ that maximally increases loss within a radius $\gamma > 0$.

Sharpness-aware minimization The most popular principal way for finding flatter minima is Sharpness-Aware Minimization (SAM) (Foret et al. 2021), which formally transforms loss minimization into a min-max optimization:

$$\min_\theta \max_{\|\Delta\|\leq\gamma} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f(x;\theta+\Delta), y)], \tag{4}$$

As formulated in the work of (Foret et al. 2021), when adding the loss value at θ, only the first term of (3) remains, yielding the objective function in (4). Through the minimization of the maximization term, SAM aims to find the minima θ with flatter loss surfaces around γ radius.

How Can Augmentations Enhance Robustness?

In this section, we provide a rigorous link between *data augmentations* and *improved generalization capability*. Our main claims are twofold: **i)** Translation between equivalent input manifold and parameter manifold, **ii)** Association of flatter loss in the input space via augmentations with parameter space flatness, leading to a reduced generalization gap.

[1] By formulating augmentation as an additive perturbation, we transform the perturbation on the parameter space to one on the input space. Details are in the following section.

Translation Between Input and Parameter Perturbations

Consider a perturbation $\delta \in \mathbb{R}^n$ around an input vector x. Intuitively, there should be a corresponding perturbation $\Delta \in \mathbb{R}^p$ around the parameter θ that yields the same output, i.e., $f(x+\delta;\theta) = f(x;\theta+\Delta)$. Moreover, for any perturbation in the closed ball $\{\delta \in \mathbb{R}^n : \|\delta\| \leq \gamma\}$ in the input space, we hypothesize an *equivalent* or *compensatory* region $\mathcal{C}_\Theta$ in the parameter space—that is, for every $\|\delta\| \leq \gamma$, there is a matching parameter perturbation $\Delta \in \mathcal{C}_\Theta$ producing the same output.

Conversely, the same intuition in reverse applies to parameter perturbations. Consider a perturbation $\Delta \in \mathbb{R}^p$ around θ. For any $\|\Delta\| \leq \gamma$, we posit an analogous *compensatory* region $\mathcal{C}_\mathcal{X}$ in the input space such that $f(x;\theta+\Delta) = f(x+\delta;\theta)$. Specifically, there is some input shift $\delta \in \mathcal{C}_\mathcal{X}$ that replicates the same output change induced by Δ.

Our goal is to formalize this duality of perturbations in both the input and parameter spaces. We now present the notion of *functional compensatory sets*, which identifies the precise sets of input shifts needed to compensate for every parameter perturbation of norm up to γ, and vice versa.

Functional Compensatory Sets

Definition 1. *(Functional Compensatory Sets)*

(A) Parameter-to-Input. *Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^N$ and a parameter perturbation set $\{\Delta \in \mathbb{R}^p : \|\Delta\| \leq \gamma\}$, a* functional compensatory set *$\mathcal{C}_\mathcal{X}$ in the input space satisfies:*

$$\text{For every } x \in \mathcal{D} \text{ and } \|\Delta\| \leq \gamma, \text{ there exists } \delta \in \mathcal{C}_\mathcal{X} \text{ s.t. } f(x;\theta+\Delta) = f(x+\delta;\theta).$$

This set $\mathcal{C}_\mathcal{X}$ is not necessarily unique, and it characterizes all possible input shifts that replicate the effect of parameter perturbations within norm γ.

(B) Input-to-Parameter. *Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^N$ and an input perturbation set $\{\delta \in \mathbb{R}^n : \|\delta\| \leq \gamma\}$, a* functional compensatory set *$\mathcal{C}_\Theta$ in the parameter space satisfies:*

$$\text{For every } x \in \mathcal{D} \text{ and } \|\delta\| \leq \gamma, \text{ there exists } \Delta \in \mathcal{C}_\Theta \text{ s.t. } f(x+\delta;\theta) = f(x;\theta+\Delta).$$

This set $\mathcal{C}_\Theta$ need not be unique, and captures all possible parameter shifts that replicate the effect of any input perturbation within norm γ.

Formalizing the Input–Parameter Perturbation Duality

The following result shows that an n-dimensional ball in the input space, whose radius depends on dataset $\mathcal{D}$ and the Jacobian's singular values, can compensate for any parameter perturbation $\|\Delta\| \leq \gamma$. Conversely, the same analysis guarantees that a p-dimensional ball in the parameter space, determined by $\mathcal{D}$ and the Jacobian, can compensate for any input perturbation $\|\delta\| \leq \gamma$.

Let $\mathbf{J}$ be the Jacobian of a function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^c$, where $\mathbb{R}^n, \mathbb{R}^p$, and $\mathbb{R}^c$ are the input, parameter, and output spaces. The Jacobian matrix $\mathbf{J} \in \mathbb{R}^{c\times(n+p)}$ can be decomposed into its input side and the parameter side partial derivatives, i.e. $\mathbf{J} = [\mathbf{J}_\mathbf{x} \ \mathbf{J}_{\boldsymbol{\theta}}]$, where $\mathbf{J}_\mathbf{x} \in \mathbb{R}^{c\times n}$ and $\mathbf{J}_{\boldsymbol{\theta}} \in \mathbb{R}^{c\times p}$. We will use the notation $\mathbf{J}_\mathbf{x}(x_i)$ and $\mathbf{J}_{\boldsymbol{\theta}}(\theta)$ when we need to explicitly represent the evaluation of $\mathbf{J}_\mathbf{x}$ and $\mathbf{J}_{\boldsymbol{\theta}}$ at point $x_i \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^p$, respectively. Finally, let $\sigma_{\min}^{x_i}$ and $\sigma_{\max}^{\theta}$ represent the mininum and the maximum values of the matrix $\mathbf{J}_\mathbf{x}(x_i)$ and $\mathbf{J}_{\boldsymbol{\theta}}(\theta)$.

Grounded on an o-minimal structure and the lazy training phenomenon of conventional neural networks, we utilize Sard's theorem (Sard 1942; Morse 1939) together with Taylor expansion to build the functional compensatory sets. We construct two sets—$\mathcal{S}_\mathcal{X}$ and $\mathcal{S}_\Theta$—that fulfill the definition of functional compensatory sets $\mathcal{C}_\mathcal{X}$ and $\mathcal{C}_\Theta$ as follows:

Theorem 1. *(Input–Parameter Duality via Functional Compensatory Sets) Consider a function $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^c$.*

(A) Parameter-to-Input *Define the ball of parameter perturbations $\{\Delta \in \mathbb{R}^p : \|\Delta\| \leq \gamma\}$ and*

$$\mathcal{S}_\mathcal{X} = \left\{\delta \in \mathbb{R}^n : \|\delta\| \leq \max_{x\in\mathcal{D}}\left(\frac{\sigma_{\max}^{\theta}}{\sigma_{\min}^{x}}\right)\gamma\right\}.$$

Then, $\mathcal{S}_\mathcal{X}$ is always a functional compensatory set for the parameter perturbation ball and dataset $\mathcal{D}$.

(B) Input-to-Parameter *Conversely, define the ball of input perturbations $\{\delta \in \mathbb{R}^n : \|\delta\| \leq \gamma\}$ and*

$$\mathcal{S}_\Theta = \left\{\Delta \in \mathbb{R}^p : \|\Delta\| \leq \max_{x\in\mathcal{D}}\left(\frac{\sigma_{\max}^{x}}{\sigma_{\min}^{\theta}}\right)\gamma\right\}.$$

Then, $\mathcal{S}_\Theta$ is always a functional compensatory set for the input perturbation ball and dataset $\mathcal{D}$.

Proof. See Appendix B.1. □

Remark 1.1. ***(Loss Stability Across Input and Parameter Spaces)*** *Ensuring the model's loss is low on all datapoints within a small neighborhood around x (i.e., for all $x+\delta$ with $\delta \in \mathcal{S}_\mathcal{X}$) guarantees that the corresponding neighborhood in parameter space $\{\theta+\Delta : \|\Delta\| \leq \gamma\}$ also has low loss. Conversely, if the model's loss is low within a small neighborhood around the parameters (i.e. $\theta+\Delta$ for $\Delta \in \mathcal{S}_\Theta$), then any corresponding input perturbation δ with $\|\delta\| \leq \gamma$ also yields low loss. Hence, stability in the input space induces stability in the parameter space, and vice versa.*

Linking Augmentations to Model Flatness and Generalization

Built upon the theorems above, we now formalize how augmentations can lead to a flatter loss landscape around the minima and the improved generalization capability. Before providing the details, let us rephrase the formal definition of flat minima, which is called b-flat minima (Shi et al. 2021):

Definition 2. *(b-flat local minima) Given loss $\mathcal{L}(\cdot,\cdot)$ and dataset $\mathcal{D}$, a model parameter $\theta \in \mathbb{R}^p$ is b-flat minima if the followings hold for the perturbation on parameter $\Delta \in \mathbb{R}^p$:*

$$\text{For all } \|\Delta\| \leq b, \ \hat{\mathcal{E}}_\mathcal{D}(\theta+\Delta) = \hat{\mathcal{E}}_\mathcal{D}(\theta) \text{ and}$$
$$\text{For some } \|\Delta\| > b, \ \hat{\mathcal{E}}_\mathcal{D}(\theta+\Delta) > \hat{\mathcal{E}}_\mathcal{D}(\theta).$$

Distance	CIFAR-10					CIFAR-100				
	AugMix	RandAug	PixMix	DeepAug	StyleAug	AugMix	RandAug	PixMix	DeepAug	StyleAug
$\gamma = 0.01$	0.0025	0.0064	0.0014	*0.0000*	*0.0000*	0.0024	0.0060	0.0011	*0.0000*	*0.0000*
$\gamma = 0.05$	**0.0100**	0.0064	0.0015	*0.0000*	*0.0000*	**0.0107**	0.0060	0.0011	*0.0000*	*0.0000*
$\gamma = 0.10$	**0.0205**	0.0065	0.0015	*0.0000*	*0.0000*	**0.0202**	0.0061	0.0012	*0.0000*	*0.0000*
$\gamma = 0.50$	**0.0959**	**0.0391**	0.0015	0.0001	*0.0000*	**0.0993**	**0.0378**	0.0012	0.0001	*0.0000*

Table 1: Empirical CDF $\big(F_N(\gamma)\big)$ for selected distance thresholds (γ) on CIFAR datasets, illustrating sample density near the original image over different augmentation methods. The higher the eCDF value, the stronger an augmentation satisfies PSA condition. Sorted in descending order. (*Italic* entries have near-zero probability $(< 5 \cdot 10^{-5})$, and **bold** entries exceeds 0.01.)

For input x, an augmentation $\mathcal{A}$ induces a distribution of variants $\tilde{x}$ near x. Let $\gamma_{\mathcal{A}}$ be the largest radius of a ball around x with nonzero probability density. If $\gamma_{\mathcal{A}} > 0$, we say $\mathcal{A}$ covers the local neighborhood. We now formalize this property and later show that it is critical for robustness.

Condition. *(Proximal-Support Augmentation (PSA))*
Given $x \in \mathbb{R}^n$ and augmentation $\mathcal{A}(\cdot)$, the probability density function $\mathcal{P}_{\mathcal{A}}(\tilde{x}|x)$ satisfies:

$$\textit{For all } \|\delta\| \leq \gamma_{\mathcal{A}},\ \mathcal{P}_{\mathcal{A}}(\tilde{x}|x) > 0, \tag{5}$$

where $\delta = \tilde{x} - x$ and $\gamma_{\mathcal{A}}$ is some positive real number.

PSA is a condition that covers any label-preserving augmentation and ties it to flat minima. It serves as a testable rule by setting a local coverage requirement around x. Discarding dependencies on the model's interpolation ability between data points, PSA is a strong condition for zero loss over all neighborhoods of inputs for any architectures.

Let $\gamma_{\mathcal{A}}$ be the radius value for PSA, and $\tilde{\mathcal{D}}$ represent the augmented dataset with respect to $\mathcal{D}$ under PSA condition. Let Θ^* and $\tilde{\Theta}^*$ be the sets of the optimal model parameters whose elements $\theta^* \in \Theta^*$ and $\tilde{\theta}^* \in \tilde{\Theta}^*$ satisfy the following equalities, $\mathcal{E}_{\mathcal{D}}(\theta^*) = 0$ and $\mathcal{E}_{\tilde{\mathcal{D}}}(\tilde{\theta}^*) = 0$, respectively. Then our claim is that the minimum b-flatness among the solution parameters in $\tilde{\Theta}^*$, shows large b (flatter) than the minimum b-flatness among the solutions in Θ^*, which is trained on $\mathcal{D}$:

Theorem 2. *(Flatness of $\tilde{\theta}^*$) Let $\theta^* \in \Theta^*$ and $\tilde{\theta}^* \in \tilde{\Theta}^*$ be b^* and $\tilde{b}^*$-flat minima, respectively. The following inequality holds:*

$$\min_{\theta^* \in \Theta^*} b^* \leq \min_{\tilde{\theta}^* \in \tilde{\Theta}^*} \tilde{b}^*.$$

Proof. See Appendix B.2. □

Remark 2.1. ***(Any augmentation with nearby representations around the original promotes flatter minima)*** *The key understanding of the theorem above is that if augmentations cover the ball-shaped region around the given original sample with radius $\gamma_{\mathcal{A}}$, then the model $\tilde{\theta}^*$ suppresses the loss values of the region. Next, the flat-region on the input space is translated to the functional compensatory set region on parameter space, which at least contains $\{\delta : \|\delta\| \leq \big(\max_{x \in \mathcal{D}} \sigma^{\theta}_{\max}/\sigma^{x}_{\min}\big)^{-1} \cdot \gamma_{\mathcal{A}}\}$.*

The final linkage to the generalization capability is straightforward by relying on the prior theoretical results that bridge Robust Risk Minimization (RRM) and the generalization bound (Cha et al. 2021). As RRM targets flat minima, Theorem 2 shows that augmentations further flatten the optimum's neighborhood, and directly tightens the bound in Theorem 3. Let us define $\gamma_{\Theta} := \big(\max_{x \in \mathcal{D}} \sigma^{\theta}_{\max}/\sigma^{x}_{\min}\big)^{-1} \cdot \gamma_{\mathcal{A}}$, then the following theorem holds:

Theorem 3. *(Generalization bound) Given M covering sets $\{\Theta_k\}_{k=1}^{M}$ of parameter space Θ with $\Theta = \bigcup_{k=1}^{M} \Theta_k$ and $diam(\Theta) = \sup_{\theta,\theta' \in \Theta} ||\theta - \theta'||_2$, where $M = \left\lceil \frac{diam(\Theta)}{\gamma_{\Theta}} \right\rceil^p$, and VC dimension v_k for each Θ_k, the following inequality holds with probability at least $1 - \delta$:*

$$\mathcal{E}_{\mathcal{T}}(\tilde{\theta}^*) < \hat{\mathcal{E}}_{\tilde{\mathcal{D}}}(\tilde{\theta}^*) + \frac{1}{2}\textbf{\textit{Div}}(\mathcal{D}, \mathcal{T}) + \max_{k \in [1,M]} \left[\sqrt{\frac{v_k \log(N/v_k)}{N} + \frac{\log(M/\delta)}{N}} \right], \tag{6}$$

where $\textbf{\textit{Div}}(\mathcal{D}, \mathcal{T}) = 2 \sup_A |\mathbb{P}_{\mathcal{D}}(A) - \mathbb{P}_{\mathcal{T}}(A)|$ measures the maximal discrepancy between the source and the target distributions $\mathcal{D}$ and $\mathcal{T}$, and N is the number of samples drawn from $\tilde{\mathcal{D}}$.

Proof. See Appendix B.3. □

Remark 3.1. ***(Augmentations can improve generalization against data distribution shifts)*** *The theorem implies that the minimization of empirical loss for the augmented dataset, i.e., $\hat{\mathcal{E}}_{\tilde{\mathcal{D}}}(\tilde{\theta}^*)$, directly aims the tighter generalization bound on target distribution $\mathcal{T}$. Also, the term M is related to $\gamma_{\mathcal{A}}$, which measures how augmented data distribution covers a wider range (referred to PSA condition.) When augmentation covers a wider range around the original sample, i.e., a larger $\gamma_{\mathcal{A}}$, it suppresses M, leading to the smaller last term of generalization bound. For trivial $\gamma_{\mathcal{A}} \simeq 0$, the flatness and bound tightness gains vanish.*

Empirical Examination on Proximal Density, Model Flatness, and Robustness

To empirically validate our theory, we examine proximal density, model flatness, and robustness on different benchmarks. Earlier, we established that any label-invariant augmentation assigning nonzero probability density to the neighborhood of the original image (PSA condition) maps to an analogous region in parameter space (Theorem 1),

Dataset	Metrics	ERM	AugMix	RandAug	PixMix	DeepAug	StyleAug
CIFAR-10	$\mu_{\text{PAC-Bayes}} \downarrow$	168.83	**117.16** (-51.67)	**110.80** (-58.03)	**102.87** (-65.96)	**67.74** (-101.09)	**122.80** (-46.03)
	LPF $\downarrow$	1.43	**0.54** (-0.89)	**0.45** (-0.98)	**0.37** (-1.06)	**0.72** (-0.71)	**1.13** (-0.30)
	$\epsilon_{\text{sharp}} \downarrow$	40.90	**25.16** (-15.74)	**24.47** (-16.43)	**26.57** (-14.33)	55.63 (+14.73)	130.64 (+89.74)
CIFAR-100	$\mu_{\text{PAC-Bayes}} \downarrow$	181.10	**155.74** (-25.36)	**149.79** (-31.31)	**134.58** (-46.52)	**72.24** (-108.86)	**129.02** (-52.08)
	LPF $\downarrow$	2.37	**1.84** (-0.53)	**1.69** (-0.68)	**1.42** (-0.95)	2.37 (0.00)	3.95 (+1.58)
	$\epsilon_{\text{sharp}} \downarrow$	42.39	**39.32** (-3.07)	**37.76** (-4.63)	**33.14** (-9.25)	84.48 (+42.09)	302.81 (+260.42)
Improvement Rate			**6/6**	**6/6**	**6/6**	3/6	2/6

Table 2: Flatness metrics of ERM and different augmentations on CIFAR datasets. Overall, when augmentation meets the PSA condition (nonzero probability near the original image) strongly, the flatter minima are found consistently. (↓*: The lower the better, i.e., flatter minimum.*)

Benchmarks	ERM	AugMix	RandAug	PixMix	DeepAug	StyleAug
CIFAR-10-C $\downarrow$	30.54	**15.24** (-15.30)	**19.65** (-10.89)	**10.60** (-19.94)	31.92 (+1.38)	**30.50** (-0.04)
CIFAR-10-$\overline{\text{C}}$ $\downarrow$	31.35	**20.28** (-11.07)	**20.64** (-10.71)	**14.60** (-16.75)	36.75 (+5.40)	36.57 (+5.22)
CIFAR-100-C $\downarrow$	59.04	**42.64** (-16.40)	**46.59** (-12.45)	**35.20** (-23.84)	62.34 (+3.30)	70.91 (+11.87)
CIFAR-100-$\overline{\text{C}}$ $\downarrow$	62.43	**48.38** (-14.05)	**48.32** (-14.11)	**40.20** (-22.23)	67.91 (+5.48)	76.56 (+14.13)
CIFAR-10, $L_2 \downarrow$	77.61	**70.76** (-6.85)	**76.15** (-1.46)	**65.81** (-11.80)	91.18 (+13.57)	91.06 (+13.45)
CIFAR-10, $L_\infty \downarrow$	98.49	99.10 (+0.61)	99.86 (+1.37)	99.46 (+0.97)	100.00 (+1.51)	99.98 (+1.49)
CIFAR-100, $L_2 \downarrow$	98.73	**92.76** (-5.97)	**96.06** (-2.67)	**90.69** (-8.04)	**98.44** (-0.29)	99.69 (+0.96)
CIFAR-100, $L_\infty \downarrow$	99.94	**99.67** (-0.27)	99.94 (0.00)	99.69 (-0.25)	99.99 (+0.05)	99.98 (+0.04)
Improvement Rate		**7/8**	**6/8**	**6/8**	1/8	1/8

Table 3: Comparison of different augmentation methods against ERM across multiple robustness scenarios over CIFAR datasets, including common corruptions (CIFAR-10/100-C/$\overline{\text{C}}$) and adversarial attacks (L_2, L_∞). In essence, augmentations that closely adhere to the PSA condition enhance model robustness in general. Conversely, augmentations that do not enhance robustness are ones that fail to satisfy the PSA condition. (↓*: The lower the better, i.e., lower error.*)

which in turn induces flatter minima (Theorem 2) and yields a tighter generalization bound (Theorem 3). Therefore, when augmentation populates rich proximal representations for a given dataset (PSA condition strongly met), it should yield flat minima and strong robustness gains under distribution shifts. In contrast, when augmentation fail to bring robustness, the PSA condition will not be satisfied.

Augmentation Methods to Be Considered

Following the taxonomy of Xu et al. (2023), augmentations can be categorized into *model-free*, *model-based*, and *policy-based* augmentations. Among existing augmentation strategies, we focus on established methods shown to improve robustness or generalization, rather than on the latest or most basic techniques. For the *model-free* augmentation type, we select **AugMix** (Hendrycks et al. 2021b) and **PixMix** (Hendrycks et al. 2022), which are augmentations relying on mixing multiple clean images to the original image. For the *model-based* augmentation type, we choose **StyleAug** (Jackson et al. 2019) and **DeepAug** (Hendrycks et al. 2021a), which utilize image-to-image models to diversify the style of the clean images. Finally, among the *policy-based* augmentation type, we consider **RandAugment** (Cubuk et al. 2020), which learns the policy for augmenting images.

Tests on Proximal Density

We first measure each augmentation's proximal density to the original image over different datasets via observing the empirical cumulative distribution function (eCDF) of L_2 distances on the widely used CIFAR and (*tiny*)ImageNet datasets (Deng et al. 2009; Krizhevsky, Hinton et al. 2009). While there exist sophisticated visual-semantic distance metrics (e.g., SSIM, LPIPS), they are not input space metrics. Thus, we have chosen L_2 distance to reflect our input-parameter space duality. Specifically, we compute how quickly the eCDF accumulates at smaller values, which provides a clear indicator of each augmentation's density near the original image. Formally, the eCDF of distances $\{d_i\}_{i=1}^N$ is defined as $F_N(\gamma) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{d_i \leq \gamma\}$, where $\mathbf{1}\{\cdot\}$ is the indicator. High values in the low-distance region of $F_N(\gamma)$ imply a large fraction of augmented samples remain close to the original image, signaling strong PSA compliance. In contrast, a low accumulation near small γ indicates weak adherence. Our analysis reveals that, for the CIFAR datasets (Table 1), AugMix, RandAug, and PixMix fulfill the PSA condition, whereas DeepAug and StyleAug do not comply, ranked in decreasing order of adherence. For (*tiny*)ImageNet, AugMix, DeepAug, PixMix, and RandAug conform to the PSA condition in descending order, with StyleAug again underperforming.

Distance	*tiny*ImageNet					ImageNet				
	AugMix	DeepAug	PixMix	RandAug	StyleAug	AugMix	DeepAug	PixMix	RandAug	StyleAug
$\gamma = 20.0$	0.0010	0.0008	0.0005	0.0003	*0.0000*	0.0005	0.0005	0.0002	0.0001	*0.0000*
$\gamma = 30.0$	0.0054	0.0047	0.0023	0.0014	*0.0000*	0.0031	0.0028	0.0013	0.0009	*0.0000*
$\gamma = 40.0$	**0.0182**	**0.0163**	0.0084	0.0051	0.0009	**0.0113**	**0.0110**	0.0054	0.0034	0.0001
$\gamma = 50.0$	**0.0446**	**0.0399**	**0.0237**	**0.0136**	0.0087	**0.0317**	**0.0311**	**0.0169**	**0.0104**	0.0008

Table 4: Empirical CDF $\big(F_N(\gamma)\big)$ for selected distance thresholds (γ) on (*tiny*)ImageNet datasets, illustrating sample density near the original image over different augmentation methods. Once again, higher eCDF values mean stronger PSA condition compliance in augmentations. Sorted in descending order.

Tests on Model Flatness

We now explore how an augmentation's density near the original image affects flatness in model parameters, linking our results from Theorem 1 and 2 with empirical flatness.

Flatness metrics. We assess each model using three quantitative measures: PAC-Bayesian flatness measure $\mu_{\text{PAC-bayes}}$ (Jiang et al. 2020), local sharpness ϵ_{sharp} (Keskar et al. 2017), and Loss-Pass Filter (LPF) metric, a strong indicator of generalization (Bisla, Wang, and Choromanska 2022). Each of these metrics reflects the difference between the proximal loss around a model parameter and its original loss, which aligns well with our theoretical framework. We therefore avoid purely "pinpoint-based" measures (e.g., maximum Hessian eigenvalues or their traces), as they do not fully capture our theorems' emphasis on local neighborhoods in parameter space. For details, see Appendix C.2.

Experimental results. We train WideResNet-40-2 on CIFAR datasets and ResNet18 on *(tiny)*ImageNet dataset, following common conventions. For details on training, see Appendix C.2. Table 2 and 5 report the flatness metrics on CIFAR and *(tiny)*ImageNet. On CIFAR, PSA compliance (Table 1 and 4) is mostly monotone with flatness. Augmentations with larger proximal density (AugMix, RandAug, PixMix) consistently beat ERM, whereas DeepAug and StyleAug give weaker gains. On *(tiny)*ImageNet, the same trend holds but is less ordered—AugMix and PixMix improve all six metrics, DeepAug and RandAug improve most, and StyleAug again brings little benefit, indicating that PSA is a strong but not exclusive factor that drives flatness.

Tests on Model Robustness

We investigate whether augmentations having rich proximal density (firmly fulfilling PSA condition) fosters robustness against different distribution shifts (Theorem 3).

Benchmarks and Metrics Robustness denotes a model's capacity to maintain performance under unforeseen data corruptions and perturbations. We employ well-established benchmarks for common corruptions (CIFAR-10/100-C/$\overline{C}$, *tiny*ImageNet-C, ImageNet-C) (Hendrycks and Dietterich 2019; Mintun, Kirillov, and Xie 2021) and adversarial robustness (untargeted PGD-L_2 and L_∞ attacks on CIFAR, *tiny*ImageNet, and ImageNet datasets) (Madry et al. 2018). The error values in Table 3 and 6 represent mean corruption errors (mCE) and adversarial errors for the common corruption and adversarial robustness benchmarks, respectively. Further details on benchmark and metric specifications are in Appendix C.3.

Experimental results. We utilize the same WideResNet-40-2 and ResNet18 models used in flatness measures to evaluate their robustness against distribution shifts (Table 3 and 6.) Fundamentally, stronger adherence to the PSA condition reveals an augmentation's effectiveness in conferring robustness to diverse distribution shifts, consistent with Theorem 3. For further experiments including varying backbone networks (Springenberg et al. 2015; Huang et al. 2017; Xie et al. 2017), refer to Appendix C.4 and D.

Discussion and Future Work

In this section, we outline the strengths, limitations, and potential directions for future research stemming from our work. The primary contributions of our study are:

- We *first* offer a series of theories to explain how augmentation can improve robustness in *flat minima* viewpoint, end-to-end fashion.
  - While prior research has explored the relationship between augmentations and robustness, to the best of our knowledge, no existing work has utilized the concept of flat minima to establish this connection.
- We *first* establish a formal connection between the family of label-invariant augmentations and robustness under *general* distribution shifts.
  - Previous studies faced limitations such as focusing solely on: i) experimental analysis, ii) particular augmentation, or iii) specific distributional shifts.
  - In contrast, our theories and analyses do not impose such restrictions on augmentations or benchmarks. We validate our findings through experiments across diverse robustness benchmarks.

Despite these contributions, our work has certain limitations. First, we do not provide the kind of extensive experimental validation across many benchmarks as done in technical reports. Second, our theorems do not directly cover label-manipulating augmentations such as Mixup (Zhang et al. 2018) and CutMix (Yun et al. 2019). Although our framework can be extended to incorporate label-mixing augmentation by setting δ toward the input mixture, this would require fixing a specific label-mixing rule and would restrict theory generality to the particular augmentation. Thus, we omit label-manipulating augmentations from our main anal-

Dataset	Metrics	ERM	AugMix	DeepAug	PixMix	RandAug	StyleAug
*tiny*ImageNet	$\mu_{\text{PAC-Bayes}} \downarrow$	136.83	**109.18** (-27.65)	**128.41** (-8.42)	**106.26** (-30.57)	**109.36** (-27.47)	**102.72** (-34.11)
	LPF $\downarrow$	5.95	**3.81** (-2.14)	**4.96** (-0.99)	**3.40** (-2.55)	**4.04** (-1.91)	**4.33** (-1.62)
	$\epsilon_{\text{sharp}} \downarrow$	15.35	**13.09** (-2.26)	25.02 (+9.67)	**10.72** (-4.63)	21.30 (+5.95)	35.40 (+20.05)
ImageNet	$\mu_{\text{PAC-Bayes}} \downarrow$	219.22	**210.63** (-8.59)	**211.41** (-7.81)	**187.89** (-31.33)	**215.58** (-3.64)	235.40 (+16.18)
	LPF $\downarrow$	8.70	**8.12** (-0.58)	**8.67** (-0.03)	**8.22** (-0.48)	8.76 (+0.06)	9.27 (+0.57)
	$\epsilon_{\text{sharp}} \downarrow$	36.80	**30.73** (-6.07)	**29.00** (-7.80)	**32.76** (-4.04)	**36.32** (-0.48)	60.06 (+23.26)
Improvement Rate			**6/6**	**5/6**	**6/6**	**4/6**	2/6

Table 5: Flatness metrics of ERM and different augmentations on (tiny)ImageNet datasets. Overall, augmentations with non-negligible density near the original image (i.e., better satisfying the PSA condition) tend to find flatter minima than ERM. This trend is clear but not perfectly monotone.

Benchmarks	ERM	AugMix	DeepAug	PixMix	RandAug	StyleAug
*tiny*ImageNet-C $\downarrow$	74.76	**63.48** (-11.28)	**56.14** (-18.62)	**61.51** (-13.25)	**66.92** (-7.84)	79.42 (+4.66)
ImageNet-C $\downarrow$	69.37	**69.34** (-0.03)	**56.68** (-12.69)	**61.51** (-7.86)	**66.88** (-2.49)	75.52 (+6.15)
*tiny*ImageNet, $L_2 \downarrow$	57.77	61.84 (+4.07)	**52.86** (-4.91)	58.38 (+0.61)	63.19 (+5.42)	83.13 (+25.36)
*tiny*ImageNet, $L_\infty \downarrow$	99.94	**99.93** (-0.01)	99.95 (+0.01)	99.98 (+0.04)	99.98 (+0.04)	100.0 (+0.06)
ImageNet, $L_2 \downarrow$	76.16	**76.06** (-0.10)	**75.37** (-0.79)	**75.44** (-0.72)	**76.09** (-0.07)	87.71 (+11.55)
ImageNet, $L_\infty \downarrow$	99.70	**99.69** (-0.01)	**99.62** (-0.08)	**99.63** (-0.07)	99.70 (+0.00)	99.92 (+0.22)
Improvement Rate		**5/6**	**5/6**	**4/6**	3/6	0/6

Table 6: Comparison of various augmentation methods against ERM across multiple robustness scenarios, including common corruptions (*tiny*ImageNet-C, ImageNet-C) and adversarial attacks (L_2, L_∞). Essentially, the closer an augmentation aligns with the PSA condition, the higher its potential to ensure robustness when confronted with differing distribution shifts.

Dataset	Metrics	AT	AT + AugMix
CIFAR-10	$\mu_{\text{PAC-Bayes}} \downarrow$	158.15	**113.78** (-44.37)
	LPF $\downarrow$	1.25	**0.58** (-0.67)
	$\epsilon_{\text{sharp}} \downarrow$	16.36	**10.29** (-6.07)
CIFAR-100	$\mu_{\text{PAC-Bayes}} \downarrow$	207.90	**161.89** (-46.01)
	LPF $\downarrow$	2.95	**2.11** (-0.84)
	$\epsilon_{\text{sharp}} \downarrow$	20.81	**10.34** (-10.47)

Table 7: Flatness evaluations of AT (FGSM, $\epsilon = 8/255$) and AT + AugMix on CIFAR-10/100.

Robustness	Benchmarks	AT	AT + AugMix
Common Corruption	CIFAR-10-C $\downarrow$	24.12	**15.39** (-8.73)
	CIFAR-10-$\overline{\text{C}}$ $\downarrow$	3.25	**1.30** (-1.95)
	CIFAR-100-C $\downarrow$	52.50	**44.03** (-8.47)
	CIFAR-100-$\overline{\text{C}}$ $\downarrow$	7.36	**3.74** (-3.62)
Adversarial Attacks	CIFAR-10, $L_2 \downarrow$	69.24	**62.86** (-6.38)
	CIFAR-10, $L_\infty \downarrow$	94.55	98.40 (+3.85)
	CIFAR-100, $L_2 \downarrow$	89.28	**88.60** (-0.68)
	CIFAR-100, $L_\infty \downarrow$	98.42	99.75 (+1.33)

Table 8: CIFAR-10/100 robustness benchmark results of AT and AT + AugMix.

ysis, leaving integration and further schemes to future work.

While not commonly done in augmentation literature, adversarial training (AT) can also be regarded as a specific type of augmentation and serve as a baseline (Addepalli, Jain et al. 2022; Li and Spratling 2023). AT improves robustness to norm-bounded attacks but stays brittle under large distribution shifts (e.g., common corruptions). We hypothesize that such a phenomenon occurs since AT has limitations in promoting large b-flat minima due to diminutive perturbations. Therefore, combining AT with augmentations that generate samples in both short-distance and long-distance range (e.g., AugMix) helps counteract its vulnerability to large shifts, by effectively boosting $\gamma_A \gg 0$ (Table 7, 8).

Finally, it is well known that models trained with flat-minima optimizers, such as Sharpness-Aware Minimization (SAM) (Foret et al. 2021) and Stochastic Weight Averaging (SWA) (Izmailov et al. 2018), generalize better under distribution shifts. However, the mechanism by which data augmentation fosters flat minima remains underexplored. We bridge this gap by introducing a simple sufficient "proximal support" condition and proving that PSA encourages flat minima. See Appendix D for further discussions on applications, flat-minima optimizers, and adversarial training.

Related Works

Data Augmentations

Existing augmentation techniques can be categorized into model-free, model-based, and policy-based approaches, following the classification proposed in Xu et al. (2023).

Model-free augmentations are further subdivided into single-image and multi-image techniques. Single-image methods encompass basic image operations like translation,

rotation, and color jitter, as well as masking strategies such as CutOut (DeVries and Taylor 2017) and Hide-and-Seek (Kumar Singh and Jae Lee 2017). These transformations have been shown to enhance model performance on target data distributions efficiently and with low overhead. Multi-image augmentations involve composing multiple operations and blending elements from different images. Among label-manipulating multi-image methods are techniques that combine pairs of distinct images and their labels, including Mixup (Zhang et al. 2018) and CutMix (Yun et al. 2019). These approaches are widely recognized for their ability to boost model performance. Subsequently, label-preserving multi-image methods like AugMix (Hendrycks et al. 2021b) and PixMix (Hendrycks et al. 2022) have demonstrated superior effectiveness in enhancing model robustness.

Model-based augmentations leverage pretrained models to produce augmented data. Several techniques employing generative models, such as CGAN (Mirza and Osindero 2014) and its variants (Douzas and Bacao 2018; Mariani et al. 2018; Ali-Gombe and Elyan 2019; Yang and Zhou 2021), are designed to mitigate data imbalance issues. Additional methods, including DeepAugment (Hendrycks et al. 2021a), ANT (Rusak et al. 2020), and StyleAug (Jackson et al. 2019), focus on improving classifier robustness to common corruptions, adversarial attacks, or domain shifts. More recently, pretrained diffusion models combined with effective prompt engineering (Islam et al. 2024; Li et al. 2024) has proven successful in enhancing performance across various vision tasks.

Policy-based augmentations focus on designing an automatic way to determine the optimal augmentation strategies by employing reinforcement learning or adversarial training. From a pioneering method, AutoAugment (Cubuk et al. 2019) that utilizes reinforcement learning for finding the best augmentation strategies, subsequent works such as Fast AA (Lim et al. 2019), Faster AA (Hataya et al. 2020), and RandAugment (Cubuk et al. 2020) aim to enhance both the efficiency of policy search and the model performance. Adversarial training-based augmentation strategies, including AdaTransform (Tang et al. 2019), Adversarial AA (Zhang et al. 2020), and AugMax (Wang et al. 2021) leverage adversarial perturbations which maximally disturbs samples to be misclassified into other labels, finally leading to improve model robustness against unseen domains.

Despite the effectiveness of augmentation methods in enhancing model performance, prior studies have focused more on their practical use rather than on understanding their theoretical impact on model robustness to data shifts.

Augmentations and Model Robustness

A number of previous works have tried to reveal the relationship between augmentation and model robustness.

Liu et al. (2024) has provided a game-theoretic perspective on augmentations and introduces a new proxy metric for measuring common corruption robustness. Zhao et al. (2020) and Volpi et al. (2018) have theoretically found out that adversarial perturbations in the latent space can simulate worst-case distributional shifts in the data. Rebuffi et al. (2021b) have empirically found out that Mixup and CutOut with model weight averaging is shown to improve adversarial robustness. Yang et al. (2020) has shown that Mixup enlarges boundary thickness, the marginal space between differently labeled samples, which is deemed to be highly correlated with adversarial robustness and common corruptions. Yin et al. (2019) interpret the augmentation-originated gains in model robustness by explaining that augmentations make deep models utilize both high and low frequency information of images so as to enhance model robustness against data corruptions. Najafi et al. (2019) and Alayrac et al. (2019) have theoretically shown that utilizing unlabeled data in training can improve adversarial robustness. Hendrycks et al. (2021a) have empirically found out that exploiting diverse augmentations together improves robustness against both adversarial and common corruptions.

Nonetheless, most of the prior interpretations of how augmentation contributes to model robustness are either confined to specific types of augmentations and robustness or empirical analysis. None of the prior works generally explain how augmentations can theoretically improve model robustness across diverse distributional shifts.

Flat Minima and Robustness

The relationship between flat minima in the loss landscape and improved generalization performance has been well-established in several seminal works (Haddouche et al. 2025; Cha et al. 2021; Zhang et al. 2024; Bian et al. 2024; Lee and Yoon 2024). Flat minima refer to regions in the optimization landscape where the loss function remains relatively stable and insensitive to small perturbations in the model parameters, which often correlate with better generalization to unseen data. For instance, Haddouche et al. (2025) establishes a rigorous theoretical link between these flat minima and superior generalization capabilities in over-parameterized neural networks, achieving this by refining the PAC-Bayes generalization bound originally proposed in Dziugaite and Roy (2017). This refinement provides a more precise quantification of how flatter regions can lead to models that perform more reliably on out-of-distribution data.

In a similar vein, Cha et al. (2021) demonstrates that flat minima contribute to tighter bounds on generalization error, supported by both theoretical analysis and practical experiments. Their approach highlights dense weight averaging as a practical and efficient heuristic for locating such minima during training, showing through empirical validation that it consistently enhances model performance across various benchmarks. Furthermore, Zhang et al. (2024) investigates how sharpness-aware minimization (SAM) relates to adversarial training (AT), demonstrating that SAM can replicate AT's behavior under the restrictive setting of Gaussian inputs and linear models. Because their theoretical analysis is confined to binary classification and Gaussian data, our theorems and the PSA condition have an edge in applying to general data distributions.

Despite these advances, prior work still leaves the open question of how *data augmentation* itself steers optimization toward flat minima. Our contribution addresses this critical gap by forging a direct connection between data augmentation strategies and the identification of flat minima, while

also demonstrating their pivotal role in bolstering model robustness against a wide array of general distribution shifts, such as those encountered in real-world scenarios involving domain variations or noisy inputs.

Other Related Works

Jiang et al. (2023) investigates how to preserve fairness under distribution shifts by reinterpreting such shifts as equivalent perturbations to both model weights and inputs, employing a transportation function. Still, their work focuses on group fairness with respect to demographic parity and equal opportunity, and demonstrates only the feasibility of this equivalence without further extensions.

Zhong, Zhu, and Yang (2024) shows in binary classification that flatter minima mitigate the boundary tilting problem (Tanay and Griffin 2016), and empirically finds these minima correlate with greater robustness to common corruptions. However, the theory remains limited, as the boundary tilting problem is closely tied to adversarial robustness and is not explored beyond binary classification.

Conclusion

Our framework connects data augmentation to robustness through flat minima viewpoint under data shifts. Augmentations meeting the PSA condition form broad, flat regions in parameter space, while others provide limited protection. Experiments on CIFAR and ImageNet substantiate our theorems, and we expect our work to inspire next-generation augmentation methods, including foundation model-based.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00459023), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program at UNIST), (No. RS-2025-25442824, AI Star Fellowship Program at UNIST), (No. RS-2025-25441996, Development of a Virtual Tactile Signal Generation Platform Technology Based on Multimodal Vision-Tactile Integrated AGI), (No. IITP-2025-RS-2022-00156361, Innovative Human Resource Development for Local Intellectualization program).

References

Addepalli, S.; Jain, S.; et al. 2022. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 1488–1501.

Alayrac, J.-B.; Uesato, J.; Huang, P.-S.; Fawzi, A.; Stanforth, R.; and Kohli, P. 2019. Are Labels Required for Improving Adversarial Robustness? In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Ali-Gombe, A.; and Elyan, E. 2019. MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361: 212–221.

Allen-Zhu, Z.; Li, Y.; and Song, Z. 2019. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning (ICML)*, 242–252. PMLR.

Bian, A.; Li, W.; Yuan, H.; Yu, C.; Wang, M.; Zhao, Z.; Lu, A.; Ji, P.; and Feng, T. 2024. Make Continual Learning Stronger via C-Flat. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 7608–7630. Curran Associates, Inc.

Bisla, D.; Wang, J.; and Choromanska, A. 2022. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 8299–8339. PMLR.

Bolte, J.; Daniilidis, A.; and Lewis, A. 2009. Tame functions are semismooth. *Mathematical Programming*, 117(1): 5–19.

Castera, C.; Bolte, J.; Févotte, C.; and Pauwels, E. 2021. An inertial Newton algorithm for deep learning. *Journal of Machine Learning Research (JMLR)*, 22(134): 1–31.

Cha, J.; Chun, S.; Lee, K.; Cho, H.-C.; Park, S.; Lee, Y.; and Park, S. 2021. SWAD: Domain Generalization by Seeking Flat Minima. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chizat, L.; Oyallon, E.; and Bach, F. 2019. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 113–123.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 702–703.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

DeVries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*.

Douzas, G.; and Bacao, F. 2018. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with applications*, 91: 464–471.

Du, S. S.; Zhai, X.; Poczos, B.; and Singh, A. 2018. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *International Conference on Learning Representations (ICLR)*.

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

Feng, R.; Zheng, K.; Huang, Y.; Zhao, D.; Jordan, M.; and Zha, Z.-J. 2022. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 33054–33065.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations (ICLR)*.

Gao, X.; Saha, R. K.; Prasad, M. R.; and Roychoudhury, A. 2020. Fuzz testing based data augmentation to improve robustness of deep neural networks. In *Proceedings of the acm/ieee 42nd international conference on software engineering*, 1147–1158.

Haddouche, M.; Viallard, P.; Şimşekli, U.; and Guedj, B. 2025. A PAC-Bayesian Link Between Generalisation and Flat Minima. In *ALT 2025-36th International Conference on Algorithmic Learning Theory*, 1–31.

Hataya, R.; Zdenek, J.; Yoshizoe, K.; and Nakayama, H. 2020. Faster autoaugment: Learning augmentation strategies using backpropagation. In *European Conference on Computer Vision (ECCV)*, 1–16. Springer.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision (ICCV)*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2021b. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Zou, A.; Mazeika, M.; Tang, L.; Li, B.; Song, D.; and Steinhardt, J. 2022. PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures. *CVPR*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 4700–4708.

Islam, K.; Zaheer, M. Z.; Mahmood, A.; and Nandakumar, K. 2024. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27621–27630.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Gordon, A. 2018. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence(UAI)*.

Jackson, P. T.; Abarghouei, A. A.; Bonner, S.; Breckon, T. P.; and Obara, B. 2019. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, volume 6, 10–11.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

Jiang, Y.; Neyshabur, B.; Mobahi, H.; Krishnan, D.; and Bengio, S. 2020. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations (ICLR)*.

Jiang, Z. S.; Han, X.; Jin, H.; Wang, G.; Chen, R.; Zou, N.; and Hu, X. 2023. Chasing fairness under distribution shift: a model weight perturbation approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations (ICLR)*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*.

Kumar Singh, K.; and Jae Lee, Y. 2017. Hide-And-Seek: Forcing a Network to Be Meticulous for Weakly-Supervised Object and Action Localization. In *ICCV*.

Lee, J.; Xiao, L.; Schoenholz, S.; Bahri, Y.; Novak, R.; Sohl-Dickstein, J.; and Pennington, J. 2019. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Lee, T.; and Yoon, S. W. 2024. Rethinking the flat minima searching in federated learning. In *Forty-first International Conference on Machine Learning (ICML)*.

Li, L.; and Spratling, M. 2023. Data Augmentation Alone Can Improve Adversarial Training. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Li, P.; Li, D.; Li, W.; Gong, S.; Fu, Y.; and Hospedales, T. M. 2021. A Simple Feature Augmentation for Domain Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8886–8895.

Li, Y.; Dong, X.; Chen, C.; Zhuang, W.; and Lyu, L. 2024. A simple background augmentation method for object detection with diffusion model. In *European Conference on Computer Vision (ECCV)*, 462–479. Springer.

Li, Y.; and Liang, Y. 2018. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

Lim, S.; Kim, I.; Kim, T.; Kim, C.; and Kim, S. 2019. Fast autoaugment. *Neural Information Processing Systems (NeurIPS)*, 32.

Liu, Z.; Zhang, J.; He, Q.; and Wang, C. 2024. Understanding Data Augmentation From A Robustness Perspective. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6760–6764. IEEE.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations (ICLR)*.

Mariani, G.; Scheidegger, F.; Istrate, R.; Bekas, C.; and Malossi, C. 2018. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*.

Mintun, E.; Kirillov, A.; and Xie, S. 2021. On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Morse, A. P. 1939. The behavior of a function on its critical set. *Annals of Mathematics*, 40(1): 62–70.

Najafi, A.; Maeda, S.-i.; Koyama, M.; and Miyato, T. 2019. Robustness to Adversarial Perturbations in Learning from Incomplete Data. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. 2021a. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*.

Rebuffi, S.-A.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. A. 2021b. Data Augmentation Can Improve Robustness. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Neural Information Processing Systems (NeurIPS)*, volume 34, 29935–29948. Curran Associates, Inc.

Rusak, E.; Schott, L.; Zimmermann, R. S.; Bitterwolf, J.; Bringmann, O.; Bethge, M.; and Brendel, W. 2020. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision (ECCV)*.

Sard, A. 1942. The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society*.

Shi, G.; Chen, J.; Zhang, W.; Zhan, L.-M.; and Wu, X.-M. 2021. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 6747–6761.

Springenberg, J.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2015. Striving for Simplicity: The All Convolutional Net. In *International Conference on Learning Representations (ICLR) workshop track*.

Tanay, T.; and Griffin, L. 2016. A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples. In *arXiv:1608.07690*.

Tang, Z.; Peng, X.; Li, T.; Zhu, Y.; and Metaxas, D. N. 2019. Adatransform: Adaptive data transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2998–3006.

Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. *Neural Information Processing Systems (NeurIPS)*, 31.

Wang, H.; Xiao, C.; Kossaifi, J.; Yu, Z.; Anandkumar, A.; and Wang, Z. 2021. AugMax: Adversarial Composition of Random Augmentations for Robust Training. In *Neural Information Processing Systems (NeurIPS)*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 1492–1500.

Xu, M.; Yoon, S.; Fuentes, A.; and Park, D. S. 2023. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recogn.*, 137(C).

Yang, H.; and Zhou, Y. 2021. Ida-gan: A novel imbalanced data augmentation gan. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 8299–8305. IEEE.

Yang, Y.; Khanna, R.; Yu, Y.; Gholami, A.; Keutzer, K.; Gonzalez, J. E.; Ramchandran, K.; and Mahoney, M. W. 2020. Boundary thickness and robustness in learning models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A Fourier Perspective on Model Robustness in Computer Vision. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *International Conference on Computer Vision (ICCV)*.

Zhang, G.; Martens, J.; and Grosse, R. B. 2019. Fast convergence of natural gradient descent for over-parameterized neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Machine Learning (ICML)*.

Zhang, X.; Wang, Q.; Zhang, J.; and Zhong, Z. 2020. Adversarial AutoAugment. In *International Conference on Learning Representations (ICLR)*.

Zhang, Y.; He, H.; Zhu, J.; Chen, H.; Wang, Y.; and Wei, Z. 2024. On the duality between sharpness-aware minimization and adversarial training. *arXiv preprint arXiv:2402.15152*.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Neural Information Processing Systems (NeurIPS)*, 33: 14435–14447.

Zhong, L.; Zhu, K.; and Yang, G. 2024. Flatter Minima of Loss Landscapes Correspond with Strong Corruption Robustness. In *International Conference on Pattern Recognition (ICPR)*, 314–328. Springer.