Bridging Symmetry and Robustness: On the Role of Equivariance in Enhancing Adversarial Robustness

Longwei Wang[1], **Ifrat Ikhtear Uddin**[1], **KC Santosh**[1†],
Chaowei Zhang[2†], **Xiao Qin**[3], **Yang Zhou**[3†]

longwei.wang@usd.edu, ifratikhtear.uddin@coyotes.usd.edu, kc.santosh@usd.edu
cwzhang@yzu.edu.cn, xqin@auburn.edu, yangzhou@auburn.edu

[1]AI Research Lab, Department of Computer Science, University of South Dakota, USA
[2]School of Information and Artificial Intelligence, Yangzhou University, Yangzhou, China
[3]Department of Computer Science and Software Engineering, Auburn University, USA

Abstract

Adversarial examples reveal critical vulnerabilities in deep neural networks by exploiting their sensitivity to imperceptible input perturbations. While adversarial training remains the predominant defense strategy, it often incurs significant computational cost and may compromise clean-data accuracy. In this work, we investigate an architectural approach to adversarial robustness by embedding group-equivariant convolutions—specifically, rotation- and scale-equivariant layers—into standard convolutional neural networks (CNNs). These layers encode symmetry priors that align model behavior with structured transformations in the input space, promoting smoother decision boundaries and greater resilience to adversarial attacks. We propose and evaluate two symmetry-aware architectures: a parallel design that processes standard and equivariant features independently before fusion, and a cascaded design that applies equivariant operations sequentially. Theoretically, we demonstrate that such models reduce hypothesis space complexity, regularize gradients, and yield tighter certified robustness bounds under the CLEVER (Cross Lipschitz Extreme Value for nEtwork Robustness) framework. Empirically, our models consistently improve adversarial robustness and generalization across CIFAR-10, CIFAR-100, and CIFAR-10C under both FGSM and PGD attacks, without requiring adversarial training. These findings underscore the potential of symmetry-enforcing architectures as efficient and principled alternatives to data augmentation-based defenses.

1 Introduction

Adversarial robustness, defined as the capacity of deep neural networks to produce consistent predictions under small and often imperceptible input perturbations, remains a fundamental and unresolved challenge in modern machine learning. Adversarial attacks exploit a model's sensitivity to small, norm-bounded input perturbations, leading to incorrect and often high-confidence predictions [1]. These perturbations typically exploit the model's reliance on spurious, non-semantic features that do not align with the true data-generating process [2]. A key contributing factor to this vulnerability is insufficient training data: when datasets are limited in size or diversity, models tend to overfit to superficial statistical patterns, such as background textures or local pixel correlations, rather than learning robust and generalizable representations [3, 4].

† Corresponding Authors.
Code available at: https://github.com/ifratmitul/Role-of-Equivariance

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Adversarial training has become a dominant approach to mitigate this vulnerability. It enhances model resilience by explicitly injecting adversarial examples into the training process, thereby guiding the model to focus on more discriminative, semantically grounded features [5, 6]. These adversarial examples expand the effective support of the training distribution, allowing models to develop wider decision margins and improved generalization to perturbed inputs [2, 7]. Despite its success, adversarial training is not without limitations: it is computationally expensive, may degrade performance on clean data, and is often specialized to the attack types seen during training. Moreover, it addresses robustness reactively by modifying data rather than proactively by redesigning the model architecture.

This motivates a fundamental question: *Can architectural priors alone improve adversarial robustness by encouraging models to align more closely with the geometric structure of data?* In this work, we explore this question through the lens of **equivariance**, the principle that model outputs should transform predictably under known input transformations. In particular, we investigate whether *embedding symmetry priors via group-equivariant convolutions* can enhance adversarial robustness in convolutional neural networks (CNNs) even in the absence of adversarial training.

Equivariance provides a principled mechanism for enforcing inductive biases that align with underlying symmetries in data. While standard CNNs are translation-equivariant by design, they are not inherently equivariant to other common transformations such as rotations and scalings. Group-equivariant convolutions generalize standard convolutions to be equivariant under larger transformation groups, such as the discrete rotation group $\mathrm{P4}$ or scale groups [8–10]. These architectures encode transformation consistency directly into the weight-sharing scheme of the network, allowing the model to process rotated or rescaled inputs without relying on data augmentation. As a result, equivariant CNNs have demonstrated improved sample efficiency, stronger generalization, and greater interpretability across domains such as medical imaging, remote sensing, and physics-informed learning [11, 12].

Despite their success in structured learning tasks, the relationship between equivariance and adversarial robustness remains underexplored. Intuitively, adversarial perturbations often introduce changes that lie off the data manifold or violate known symmetries. By constraining the model to respond consistently along group-induced orbits and suppressing sensitivity to off-orbit perturbations, equivariant architectures may provide a natural defense mechanism. This raises a key research question: *How does architectural equivariance influence a model's resilience to adversarial perturbations, both theoretically and empirically?*

In this paper, we bridge the gap between symmetry enforcement and adversarial robustness by conducting a systematic study of CNN architectures that integrate standard, rotation-equivariant, and scale-equivariant convolutions. We propose two model designs to incorporate equivariant layers and evaluate their robustness properties across a spectrum of adversarial and natural corruption settings. Our main contributions are summarized as follows:

- We present a theoretical analysis demonstrating that equivariant architectures contract the hypothesis space, regularize gradient behavior, and admit tighter certified robustness bounds under the CLEVER (Cross Lipschitz Extreme Value for nEtwork Robustness) framework.
- We propose and compare two symmetry-aware CNN architectures *parallel* and *cascaded* that integrate standard, rotation-equivariant, and scale-equivariant convolutional layers. We show that the parallel design better preserves complementary feature spaces and achieves superior robustness. We explore two fusion strategies *simple concatenation* and *weighted summation* for combining features from multiple symmetry branches. Our findings indicate that concatenation consistently outperforms weighted fusion in adversarial settings.
- We validate our approach through extensive experiments on CIFAR-10, CIFAR-10C, and CIFAR-100 datasets, using FGSM and PGD attacks. Our results show that equivariant CNNs, particularly the parallel design with combined rotation and scale branches, significantly outperform standard CNNs in adversarial accuracy without requiring adversarial training.

2 Related Works

Trustworthy machine learning, which focuses on developing and deploying machine learning models that are not only accurate but also robust, private, fair, and explainable, has attracted active research in recent years [13–86]. Adversarial robustness, a core pillar of trustworthy ML, addresses the vul-

nerability of neural networks to imperceptible perturbations. In this work, we explore the intersection of symmetry-aware architectures and adversarial robustness, drawing from two key research areas.

2.1 Equivariant Neural Networks

Equivariance in neural networks ensures that transformations applied to input data lead to predictable and consistent transformations in the learned representations, aligning models with the inherent symmetries in the data. A landmark advancement in this field is the introduction of Group Equivariant Convolutional Networks (G-CNNs) [8], which extended traditional convolutional operations to group transformations, including rotations [87][88]. These networks demonstrated significant gains in performance and efficiency on symmetry-rich datasets such as MNIST and CIFAR-10. Subsequent progress led to the development of Harmonic Networks, which employed circular harmonics to achieve continuous rotational equivariance [9]. By eliminating the need for discrete approximations, these models improved the flexibility of equivariant frameworks [89]. Further extending these ideas, Scale-Equivariant Steerable Networks were introduced to address scale transformations, enabling the processing of multi-scale inputs without explicit data augmentation [90]. Steerable CNNs provided a versatile framework for handling equivariance under a range of transformations, paving the way for applications in domains such as medical imaging, astrophysics, and 3D object recognition [91].

More recently, spherical equivariance has garnered attention, with spherical CNNs being developed to handle data defined on spherical domains [12][92][93]. These models have found use in global-context tasks, including climate modeling and astrophysics [11]. Additionally, equivariant networks have been applied in molecular biology, utilizing molecular symmetries to predict chemical properties [94]. Despite these advancements, the susceptibility of equivariant models to adversarial perturbations remains an open area of investigation.

2.2 Adversarial Robustness

The discovery of adversarial examples exposed a significant limitation in neural networks, revealing their vulnerability to small, carefully designed perturbations [1]. These adversarial inputs exploit weaknesses in CNN architectures, leading to incorrect predictions. The Fast Gradient Sign Method (FGSM) formalized this issue as a single-step attack based on the direction of the loss gradient [95]. Later, Projected Gradient Descent (PGD) was introduced as a stronger iterative attack method, becoming a benchmark for adversarial testing [5]. Efforts to defend against such attacks have primarily focused on adversarial training, where models are trained using adversarially perturbed data to improve robustness [5][96][97]. However, this approach often results in reduced accuracy on clean data [2]. To mitigate these trade-offs, researchers have proposed architectural innovations, such as feature denoising modules [98] and preprocessing techniques like compression, resizing, and randomization [99][100][101], which aim to diminish the effect of adversarial perturbations.

Advanced defenses have also leveraged model uncertainty and interpretability. Randomized smoothing has emerged as a certified defense strategy [102][103][104], while ensemble methods have demonstrated improved robustness by combining multiple decision boundaries [105] [106]. Despite these promising developments, the potential to integrate the symmetry-preserving principles of equivariant networks into adversarial defense strategies remains largely untapped, leaving a valuable avenue for future research.

3 Equivariance and Adversarial Robustness

Neural networks are known to be vulnerable to *adversarial perturbations*: small, human-imperceptible modifications to input data that cause incorrect predictions with high confidence. This phenomenon is often attributed to the model's overreliance on non-semantic features and irregular decision boundaries. One principled approach to mitigating this sensitivity is to enforce architectural inductive biases aligned with known symmetries of the data distribution. Among such biases, *equivariance* has emerged as a theoretically grounded and empirically effective mechanism.

3.1 Equivariance in Neural Networks

Definition 1 (Equivariant Function)**.** *Let G act on $\mathcal{X}$ via T_g and on $\mathcal{Y} \subseteq \mathbb{R}^k$ via a representation $\rho(g) \in \mathrm{Aut}(\mathcal{Y})$. A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be G-*equivariant *if:*

$$f(T_g x) = T_g f(x), \quad \forall g \in G,\, x \in \mathcal{X}.$$

Standard CNNs are translation-equivariant due to weight sharing across spatial positions. However, they lack equivariance to transformations like rotation and scaling. Group-equivariant CNNs (G-CNNs) generalize convolution to act equivariantly under more general groups G, such as C_n, $\mathrm{SO}(2)$, or dilation groups, thereby promoting symmetry-aligned representations.

Formally, let G be a group with an associated action on the input space $\mathcal{X} \subset \mathbb{R}^d$, and let $\rho : G \to \mathrm{Aut}(\mathbb{R}^k)$ be a linear representation of G acting on the feature space. A function $f : \mathcal{X} \to \mathbb{R}^k$ is said to be *equivariant* with respect to the group G if it satisfies:

$$f(g \cdot x) = \rho(g) f(x), \quad \forall g \in G, \tag{1}$$

where $g \cdot x$ denotes the transformed input under the action of $g \in G$. Equivariance ensures that applying a transformation to the input leads to a predictable transformation in the output, thereby promoting stability and consistency in feature representations.

3.2 Adversarial Robustness and Margin Bounds

Definition 2 (Adversarial Robustness)**.** *A classifier $f : \mathbb{R}^d \to \mathbb{R}^k$ is said to be (ε, p)-robust at input $x \in \mathbb{R}^d$ if:*

$$f(x + \delta) = f(x), \quad \forall \delta \in \mathbb{R}^d,\ \|\delta\|_p \le \varepsilon.$$

Definition 3 (Margin Function)**.** *Let $f_c(x)$ denote the logit score of the predicted class c, and $f_j(x)$ the score of class $j \neq c$. The class margin is:*

$$g_{c,j}(x) := f_c(x) - f_j(x).$$

Definition 4 (Certified Robustness via Lipschitz Bound)**.** *Let $g_{c,j}$ be locally Lipschitz with constant $L > 0$ near x. Then for any $\delta \in \mathbb{R}^d$ such that $\|\delta\|_p \le \varepsilon$, we have:*

$$|g_{c,j}(x + \delta) - g_{c,j}(x)| \le L \|\delta\|_p.$$

Consequently, robustness against class j is certified if:

$$\varepsilon^{(p)}_{c \to j} \ge \frac{g_{c,j}(x)}{L}.$$

Definition 5 (CLEVER Bound [107])**.** *Let $g_{c,j}$ be the margin function as above. The CLEVER bound estimates a certified perturbation radius as:*

$$\epsilon^{(p)}_{\min}(x) := \min_{j \neq c} \frac{g_{c,j}(x)}{L^{(j)}_q},$$

where $1/p + 1/q = 1$, and

$$L^{(j)}_q := \sup_{x' \in B_p(x,r)} \|\nabla g_{c,j}(x')\|_q$$

is a data-dependent estimate of the local Lipschitz constant of $g_{c,j}$ in the dual norm.

4 Theoretical Analysis of Adversarial Robustness with Equivariant Convolutions

This section presents a comprehensive theoretical framework for analyzing the adversarial robustness of group-equivariant neural networks. We develop the mathematical foundations necessary for understanding the relationship between equivariance and model sensitivity, formalize certified robustness bounds under Lipschitz constraints, and show how equivariant architectures induce smoother gradients and larger certified margins.

4.1 Mathematical Preliminaries and Equivariant Structures

Definition 6 (Input Space and Model)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space, and let $f : \mathcal{X} \to \mathbb{R}^k$ be a neural network mapping inputs to logit vectors. We assume f is differentiable almost everywhere.*

Definition 7 (Orbit and Quotient Space)**.** *The orbit of a point $x \in \mathcal{X}$ under G is the set $[x]_G := \{g \cdot x \mid g \in G\}$. The quotient space $\mathcal{X}/G$ is the collection of distinct orbits in $\mathcal{X}$.*

Definition 8 (Jacobian and Lipschitz Constant)**.** *If f is differentiable at x, the Jacobian is $J_f(x) := \nabla f(x) \in \mathbb{R}^{k \times d}$, and the local Lipschitz constant is $L(x) := \|J_f(x)\|_2$.*

Definition 9 (Adversarial Perturbation)**.** *A vector $\delta \in \mathbb{R}^d$ is an adversarial perturbation at x if $f(x + \delta) \neq f(x)$ and $\|\delta\|_p \leq \varepsilon$.*

4.2 Jacobian Structure and Lipschitz Regularity

Definition 10 (Global Lipschitz Continuity)**.** *A function f is globally Lipschitz if there exists $L > 0$ such that:*

$$\|f(x_1) - f(x_2)\| \leq L \cdot \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathbb{R}^d.$$

Definition 11 (Jacobian under Equivariance)**.** *If $f(g \cdot x) = \rho(g) f(x)$, then the Jacobian transforms as:*

$$J_f(g \cdot x) = \rho(g) J_f(x) Dg^{-1},$$

where Dg^{-1} is the Jacobian of the inverse transformation.

Lemma 1 (Jacobian Norm Invariance [108])**.** *If $\rho(g)$ and Dg^{-1} are orthogonal matrices, then:*

$$\|J_f(g \cdot x)\|_2 = \|J_f(x)\|_2.$$

4.3 Certified Robustness via CLEVER Bounds

We analyze the certified adversarial robustness of group-equivariant networks through the CLEVER framework [107], which provides a lower bound on the minimum input perturbation required to induce misclassification. We show that equivariance yields gradient invariance over group orbits, leading to consistent and stronger robustness certification.

Lemma 2 (Transformation of Margins under Group Equivariance [109])**.** *Let f be G-equivariant, i.e.,*

$$f(g \cdot \mathbf{x}) = \rho(g) f(\mathbf{x}),$$

where $\rho : G \to \mathrm{GL}(\mathbb{R}^k)$ is a linear representation. Then for any $j \neq c$,

$$g_{c,j}(g \cdot \mathbf{x}) = \rho_{cc}(g) f_c(\mathbf{x}) - \rho_{jj}(g) f_j(\mathbf{x}).$$

Lemma 3 (Gradient Transformation of Margin Function [108])**.** *Differentiating the margin function under the group action yields:*

$$\nabla g_{c,j}(g \cdot \mathbf{x}) = \rho(g) \nabla g_{c,j}(\mathbf{x}) Dg^{-1},$$

where $Dg^{-1} \in \mathbb{R}^{d \times d}$ is the Jacobian of the inverse group action.

Theorem 1 (Orbit-Invariance of Margin Gradient Norm)**.** *If both $\rho(g)$ and Dg^{-1} are norm-preserving (e.g., orthogonal matrices), then for all $g \in G$,*

$$\|\nabla g_{c,j}(g \cdot \mathbf{x})\|_q = \|\nabla g_{c,j}(\mathbf{x})\|_q.$$

As a result, the Lipschitz constant of the margin function is invariant across the group orbit:

$$L_q^{(j)} = \sup_{\mathbf{x}' \in B_p([\mathbf{x}]_G, r)} \|\nabla g_{c,j}(\mathbf{x}')\|_q,$$

where $[\mathbf{x}]_G := \{g \cdot \mathbf{x} \mid g \in G\}$.

The orbit-invariance of both the classification margin and its gradient norm establishes a compelling theoretical foundation for the robustness of group-equivariant networks. These models are not merely robust at isolated input points but offer uniform guarantees across entire equivalence classes of inputs linked by symmetry transformations. By construction, group-equivariant architectures preserve margin values consistently under group actions, ensuring that the discriminative separation between classes remains stable across symmetrically transformed instances. Furthermore, they inherently suppress gradient sensitivity in directions aligned with the symmetry structure of the data, effectively filtering out perturbations that respect these invariances. As a result, equivariant networks exhibit tighter and more reliable CLEVER-certified robustness bounds throughout the input space. Detailed proof is provided in Appendix A.1.

4.4 Equivariance-Induced Gradient Smoothing

A core source of adversarial vulnerability in neural networks is the irregularity of their input-output mappings, often reflected in sharp or non-smooth gradients. Group-equivariant networks mitigate this by imposing geometric constraints that smooth gradients over symmetric input transformations. The idea of this section is to analyze how group equivariance suppress adversarial vulnerability by smoothing the gradients of the network with respect to its inputs. Specifically, we show that group symmetries induce consistent, low-variance gradient fields along symmetric transformations, and reduce sensitivity to adversarial perturbations that deviate from these structured directions. A key source of adversarial fragility in neural networks stems from the irregularity of their input-output mappings, often manifesting as sharp or non-smooth gradients. Group-equivariant models mitigate this issue by embedding geometric constraints that align the model's behavior with inherent symmetries in the data, leading to smoother gradients and more stable decision boundaries.

Definition 12 (Logit Gradient and Jacobian Matrix)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be a differentiable classifier, and let $f_j(\mathbf{x})$ denote the logit for class j. The Jacobian of f at $\mathbf{x}$ is:*

$$J_f(\mathbf{x}) := \nabla f(\mathbf{x}) = \begin{bmatrix} \nabla f_1(\mathbf{x})^\top \\ \vdots \\ \nabla f_k(\mathbf{x})^\top \end{bmatrix} \in \mathbb{R}^{k \times d},$$

where $\nabla f_j(\mathbf{x}) \in \mathbb{R}^d$ denotes the gradient of the j-th logit with respect to the input. Each row of $J_f(\mathbf{x})$ characterizes how sensitive a particular output is to infinitesimal changes in different input directions.

Lemma 4 (Gradient Transformation under Group Equivariance [110])**.** *Let f be a G-equivariant function, i.e.,*

$$f(g \cdot \mathbf{x}) = \rho(g) f(\mathbf{x}),$$

with $\rho(g)$ a representation and Dg^{-1} the Jacobian of the inverse group action. Then:

$$\nabla f(g \cdot \mathbf{x}) = \rho(g) \cdot \nabla f(\mathbf{x}) \cdot Dg^{-1}.$$

Definition 13 (Orbit-Averaged Gradient Field)**.** *Define the per-logit gradient vector as $\boldsymbol{\phi}_j(\mathbf{x}) := \nabla f_j(\mathbf{x})$. Then, the orbit-averaged gradient is:*

$$\overline{\boldsymbol{\phi}}_j(\mathbf{x}) := \frac{1}{|G|} \sum_{g \in G} \nabla f_j(g \cdot \mathbf{x}) = \frac{1}{|G|} \sum_{g \in G} \rho(g) \nabla f_j(\mathbf{x}) Dg^{-1}.$$

Lemma 5 (Smoothing via Orbit Averaging [109])**.** *Let $\boldsymbol{\delta} \in \mathbb{R}^d$ be a small perturbation. Then:*

$$\|\overline{\boldsymbol{\phi}}_j(\mathbf{x}) - \boldsymbol{\phi}_j(\mathbf{x})\| \ll \|\boldsymbol{\phi}_j(\mathbf{x} + \boldsymbol{\delta}) - \boldsymbol{\phi}_j(\mathbf{x})\|,$$

especially when $\boldsymbol{\delta}$ is orthogonal to the group orbit $[\mathbf{x}]_G$. Thus, orbit-averaging suppresses high-frequency variations in the gradient field.

Theorem 2 (Directional Suppression of Off-Orbit Perturbations)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be a differentiable function that is equivariant under the action of a compact group G, i.e.,*

$$f(g \cdot \mathbf{x}) = \rho(g) f(\mathbf{x}) \quad \textit{for all } g \in G,$$

where $\rho(g) \in \mathrm{GL}(\mathbb{R}^k)$ is a linear representation and $g \cdot \mathbf{x}$ is the group action on the input space. Suppose a perturbation vector $\boldsymbol{\delta} \in \mathbb{R}^d$ can be decomposed as:

$$\boldsymbol{\delta} = \boldsymbol{\delta}_G + \boldsymbol{\delta}_\perp,$$

where $\boldsymbol{\delta}_G \in T_{\mathbf{x}}([\mathbf{x}]_G)$ lies in the tangent space of the group orbit at $\mathbf{x}$, and $\boldsymbol{\delta}_\perp \perp T_{\mathbf{x}}([\mathbf{x}]_G)$. Then:

$$\|\nabla f(\mathbf{x} + \boldsymbol{\delta}_\perp) - \nabla f(\mathbf{x})\|_2 \gg \|\nabla f(\mathbf{x} + \boldsymbol{\delta}_G) - \nabla f(\mathbf{x})\|_2 .$$

In particular, if f is orbit-averaged over G, then $\|\nabla f(\mathbf{x} + \boldsymbol{\delta}_G) - \nabla f(\mathbf{x})\|_2 \to 0$, and off-orbit perturbations dominate gradient variability.

This theorem formalizes the observation that equivariant networks inherently suppress gradient sensitivity along symmetry-respecting directions, while remaining susceptible to orthogonal, adversarial ones. This anisotropy in gradient variability contributes to improved adversarial robustness and smoother decision boundaries. Equivariance not only constrains functional outputs under transformations but also regularizes local geometry of the model's decision surface. Gradient smoothing and directional suppression enhance robustness by reducing sensitivity to adversarial perturbations, especially those orthogonal to the data manifold's symmetric structure. Detailed proof is provided in Appendix A.2.

4.5 Robustness Analysis of Scale Equivariance

Scale-equivariant neural networks do not satisfy the same assumptions required for certified robustness under isometric transformations such as rotations. Specifically, scale transformations alter the norm of the input, violating the orthogonality condition required in Lemma 1 and Theorem 1. Nevertheless, scale-equivariant architectures contribute to adversarial robustness through a different mechanism namely, gradient smoothing via multi-scale orbit averaging.

Let $x \in \mathbb{R}^d$ be an input and G_s a finite group of scaling transformations. The orbit of x under G_s is defined as:

$$\mathcal{O}_s(x) = \{T_s(x) \mid s \in G_s\},$$

where $T_s(x) = s \cdot x$. We define the orbit-averaged gradient of a class logit function ϕ_j as:

$$\bar{\nabla}\phi_j(x) = \frac{1}{|G_s|} \sum_{s \in G_s} \nabla\phi_j(T_s x).$$

Unlike in the rotation-equivariant case, the norms $\|\nabla\phi_j(T_s x)\|$ are not preserved across the orbit, but the averaging process acts as a form of regularization. It reduces the gradient variance across local neighborhoods, thereby smoothing the decision boundary and dampening sensitivity to adversarial perturbations that rely on sharp gradients.

Scale-equivariant convolutional neural networks (CNNs) achieve robustness by explicitly enforcing equivariance across multiple spatial scales. This is typically done using scale-group convolutions, defined as:

$$[\Phi f](x) = \bigoplus_{s \in G_s} \psi_s * f(T_s^{-1} x),$$

where ψ_s is the filter bank corresponding to scale s, $*$ denotes convolution, and $T_s(x) = s \cdot x$. The output is a scale-indexed feature map that captures the input structure across different resolutions.

This structure induces a smoothing effect both in the feature and gradient spaces. Specifically, consider the aggregated feature response at a given layer:

$$h(x) = \sum_{s \in G_s} w_s \cdot \phi_s(x), \quad \text{with } \phi_s(x) = \psi_s * f(T_s^{-1} x),$$

and its gradient with respect to the input:

$$\nabla h(x) = \sum_{s \in G_s} w_s \cdot \nabla\phi_s(x).$$

Although the gradient norms $\|\nabla\phi_s(x)\|$ scale with s, their aggregation smooths the overall gradient field by suppressing high-frequency components. This is analogous to a low-pass filter in the frequency domain and reduces the model's vulnerability to adversarial perturbations that exploit sharp local gradient changes. While scale-equivariant models fall outside the domain of certified robustness guarantees derived under norm-preserving assumptions, they introduce robustness via a complementary mechanism: smoothing the activation and gradient fields through multi-scale aggregation. This mechanism stabilizes the model's output under input perturbations and effectively regularizes its sensitivity, promoting robustness in practice.

5 Equivariance Enhanced Architectural Designs

5.1 Group Equivariant Convolutions

Group Equivariant Convolutional Networks (G-CNNs) generalize standard convolutional architectures by incorporating symmetry priors directly into the model design [8]. These networks are constructed to preserve equivariance under transformations defined by a group G, such as translations, rotations, or scalings. To achieve this, the conventional convolution operation is replaced by a *group convolution*, which aggregates features across group-transformed versions of both the input and the filters. In this work, we focus on two widely applicable instances: rotation-equivariant and scale-equivariant convolutions. Detailed formulations of these operations are provided in Appendix C.1.

5.2 Equivariance-Enforced Architectural Designs

We investigate two architectural strategies that integrate standard, rotation-equivariant, and scale-equivariant convolutions to enhance robust feature extraction: the *parallel* design and the *cascaded* design. Each approach offers a distinct balance between representational diversity and computational efficiency. Comprehensive architectural details are provided in Appendix C.2.

6 Experiments and Discussion

This section details the experimental setup used to evaluate the impact of adding rotation- and scale-equivariant convolutions on adversarial robustness and generalization. The experiments were conducted using three widely recognized datasets CIFAR-10, CIFAR-100, and CIFAR-10C to ensure a comprehensive evaluation of adversarial robustness, and generalization under natural corruptions. CIFAR-10C [111] is a variant of CIFAR-10 designed to evaluate corruption robustness. It includes 19 types of natural corruptions (e.g., Gaussian noise, motion blur, fog, and pixelation) applied at five levels of severity.

6.1 Models

To investigate the impact of equivariance on adversarial robustness, we designed and evaluated five CNN architectures with varying symmetry-aware modifications.

Baseline Standard CNN serves as the benchmark model, implemented with either 4 or 10 convolutional layers. **Parallel GCNN** replaces the first convolutional layer with two parallel branches: a standard convolution branch and a rotation-equivariant branch based on the discrete group $\mathrm{P4}$. **Parallel GCNN with Rotation- and Scale-Equivariant Branches** extends the above by introducing a third scale-equivariant branch, enabling the model to process inputs across multiple geometric transformations. **Cascaded GCNN** adopts a sequential structure, where the input is first processed by a rotation-equivariant layer, followed by standard convolutions. **Weighted Parallel GCNN** uses the same three-branch structure as the previous parallel design but replaces feature concatenation with learnable fusion weights optimized during training.

6.2 Comparison of Adversarial Robustness under FGSM and PGD Attacks

We evaluated five models Baseline Standard CNN, Parallel GCNN, Parallel GCNN with Rotation- and Scale-Equivariant Branch, Cascaded GCNN, and Weighted Parallel GCNN on CIFAR-10 and CIFAR-100 datasets under FGSM and PGD attacks. To ensure a comprehensive understanding of the impact of network depth on robustness, we experimented with both 4-layer and 10-layer CNN architectures for all models.

In Figure 1, we present the adversarial robustness comparison of five models using 4-layer architectures on CIFAR-10 and CIFAR-100. The evaluation considers adversarial accuracies under FGSM and PGD attacks across a range of perturbation magnitudes (ϵ). For CIFAR-10, the Parallel GCNN with Rotation and Scale Branch exhibited the highest robustness. The Parallel GCNN also performed well, but its robustness declined more rapidly compared to the combined model. On CIFAR-100, the overall robustness was lower due to increased class diversity and complexity. The Parallel GCNN with Rotation- and Scale-Equivariant Branch remained the most robust particularly at higher perturbations. The Parallel GCNN showed competitive performance at low perturbations but lagged behind the combined model, particularly at higher perturbations.

In Figure 2, we consider 10-layer architectures on CIFAR-10 and CIFAR-100 datasets. On both CIFAR-10 and CIFAR-100, the Parallel GCNN with Rotation and Scale-Equivariant Branch maintained the highest adversarial robustness across all perturbation levels. The Parallel GCNN with Rotation-Equivariant Branch also demonstrated strong robustness, though it was consistently outperformed by the combined model. Both the Cascaded GCNN and Weighted Parallel GCNN showed limited robustness, with adversarial accuracies dropping below 15% for FGSM and almost negligible for PGD attacks at higher perturbation levels.

To validate our theoretical framework under strict symmetry constraints, we evaluated fully equivariant architectures where all convolutional layers are equivariant, without any standard convolution branches. The 10-layer fully equivariant model achieves 73.01% FGSM and 64.96% PGD accuracy

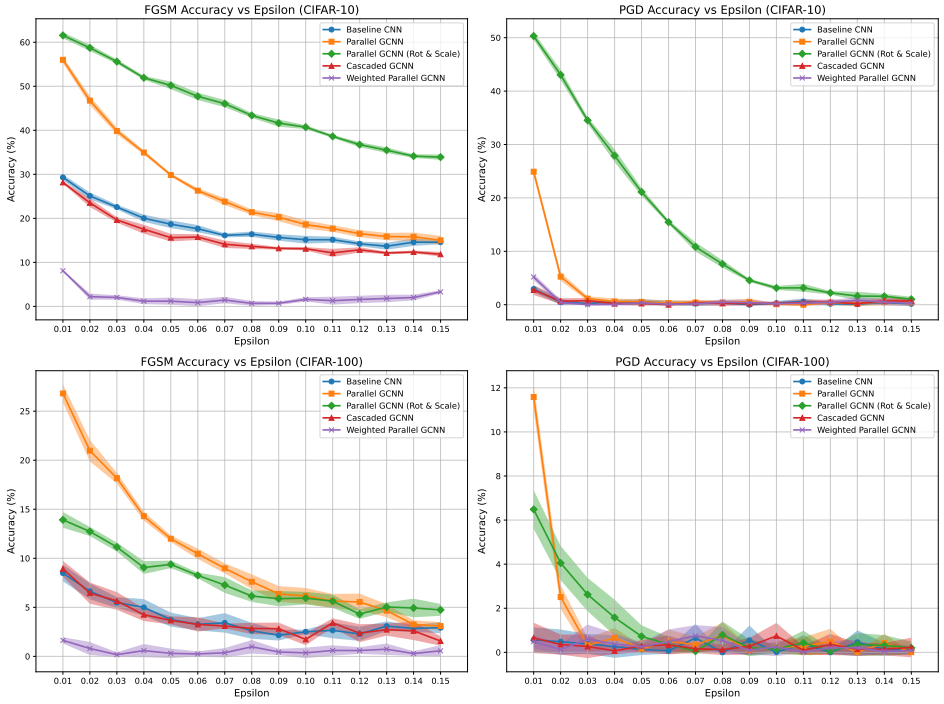

Figure 1: Adversarial robustness comparison of five models using 4-layer architectures on CIFAR-10 and CIFAR-100. Shaded regions represent ±1 standard deviation over 5 random seeds.

Table 1: Performance Analysis of Different Models under Various Corruption Types and Perturbation Levels on CIFAR10C (%)

	BaselineCNN				Cascaded GCNN				Parallel GCNN				Parallel GCNN (R & S)				Weighted GCNN (R & S)			
Corruption Type	ε_1	ε_2	ε_3	ε_4	ε_1	ε_2	ε_3	ε_4	ε_1	ε_2	ε_3	ε_4	ε_1	ε_2	ε_3	ε_4	ε_1	ε_2	ε_3	ε_4
Brightness	14.79	2.92	0.64	0.16	8.27	0.46	0.02	0.00	**18.38**	5.86	1.85	0.65	8.38	0.24	0.00	0.00	8.13	0.37	0.01	0.00
Contrast	3.25	0.71	0.39	0.33	1.31	0.00	0.00	0.00	**7.35**	2.52	0.46	0.05	0.69	0.00	0.00	0.00	0.59	0.00	0.00	0.00
Defocus Blur	7.92	1.33	0.32	0.11	3.09	0.04	0.00	0.00	**13.15**	3.22	0.79	0.20	2.72	0.03	0.00	0.00	2.80	0.04	0.00	0.00
Elastic Transform	6.80	1.11	0.28	0.07	1.90	0.03	0.00	0.00	**11.46**	2.68	0.66	0.19	1.92	0.01	0.00	0.00	2.02	0.03	0.00	0.00
Fog	3.64	0.45	0.12	0.03	1.31	0.01	0.00	0.00	**6.97**	1.24	0.21	0.04	1.39	0.00	0.00	0.00	1.11	0.00	0.00	0.00
Frost	9.32	1.45	0.30	0.10	3.15	0.17	0.01	0.00	**11.90**	3.32	1.02	0.38	3.44	0.10	0.01	0.00	3.07	0.15	0.01	0.00
Gaussian Blur	7.24	1.12	0.27	0.10	2.74	0.03	0.00	0.00	**11.82**	2.86	0.69	0.17	2.02	0.02	0.00	0.00	2.23	0.04	0.00	0.00
Gaussian Noise	5.38	1.02	0.34	0.14	1.82	0.03	0.00	0.00	**14.68**	2.66	0.53	0.12	1.84	0.00	0.00	0.00	1.78	0.01	0.00	0.00
Glass Blur	5.51	1.07	0.33	0.13	0.84	0.01	0.00	0.00	**10.70**	2.29	0.48	0.11	0.90	0.02	0.00	0.00	1.22	0.02	0.00	0.00
Impulse Noise	5.99	1.06	0.31	0.14	1.57	0.03	0.00	0.00	**12.00**	2.03	0.34	0.06	1.56	0.01	0.00	0.00	1.44	0.01	0.00	0.00
JPEG Compression	8.68	1.69	0.46	0.15	3.42	0.11	0.00	0.00	**17.44**	4.42	1.09	0.31	4.35	0.06	0.00	0.00	3.35	0.06	0.00	0.00
Motion Blur	6.34	0.91	0.23	0.07	2.11	0.02	0.00	0.00	**10.46**	2.68	0.69	0.19	1.82	0.02	0.00	0.00	1.84	0.03	0.00	0.00
Pixelate	9.04	1.66	0.40	0.14	4.45	0.14	0.00	0.00	**17.05**	4.43	1.08	0.28	5.70	0.11	0.00	0.00	4.69	0.09	0.00	0.00
Saturate	9.74	2.33	0.79	0.29	8.18	0.44	0.02	0.00	**22.09**	6.51	1.70	0.53	7.90	0.20	0.01	0.00	7.82	0.29	0.01	0.00
Shot Noise	5.62	1.07	0.38	0.17	2.36	0.05	0.00	0.00	**16.03**	3.09	0.61	0.13	2.71	0.01	0.00	0.00	2.33	0.03	0.00	0.00
Snow	9.71	1.85	0.52	0.19	3.71	0.22	0.02	0.00	**15.06**	3.95	1.23	0.42	4.08	0.14	0.02	0.00	3.93	0.19	0.03	0.01
Spatter	7.80	1.58	0.49	0.23	3.19	0.11	0.01	0.00	**15.41**	3.27	0.77	0.19	3.54	0.05	0.00	0.00	3.29	0.07	0.01	0.00
Speckle Noise	5.25	1.15	0.38	0.18	2.48	0.04	0.00	0.00	**15.81**	2.91	0.53	0.10	2.92	0.01	0.00	0.00	2.24	0.02	0.00	0.00
Zoom Blur	6.39	0.93	0.20	0.08	1.47	0.00	0.00	0.00	**10.41**	2.34	0.59	0.14	1.07	0.01	0.00	0.00	1.39	0.03	0.00	0.00
Mean	7.28	1.34	0.38	0.15	2.76	0.10	0.00	0.00	**13.59**	3.28	0.81	0.23	2.95	0.05	0.00	0.00	3.01	0.07	0.00	0.00
Std	±2.63	±0.58	±0.16	±0.07	±1.64	±0.13	±0.00	±0.00	**±3.82**	±1.29	±0.43	±0.16	±2.20	±0.07	±0.00	±0.00	±2.01	±0.10	±0.00	±0.00

[†] **R & S** represents Rotation and Scaling. ε_i represents perturbation threshold where $i \in \{1, \ldots, 4\}$ with values $\{0.01, 0.02, 0.03, 0.04\}$ respectively.
Bold: Best performance across all models. Results averaged over 5 runs with standard deviation shown.

at $\varepsilon = 0.01$ on CIFAR-10, confirming that orbit-invariant gradient regularization compounds beneficially when symmetry is enforced end-to-end. Complete results for fully equivariant architectures are provided in Appendix D.

Our equivariant models achieve these robustness improvements without adversarial training. For context, we compare our 10-layer equivariant model against a standard CNN trained with PGD adversarial training in Appendix E.

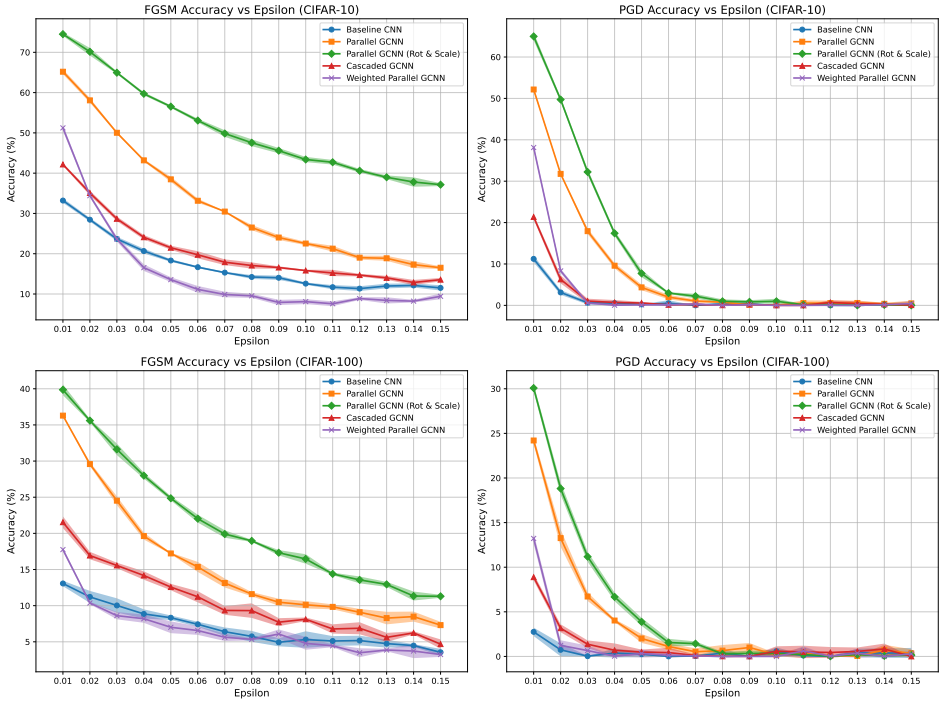

Figure 2: Adversarial robustness comparison of five models using 10-layer architectures on CIFAR-10 and CIFAR-100. Shaded regions represent ±1 standard deviation over 5 random seeds.

Table 1 presents the performance analysis of various models on CIFAR-10C under a range of corruption types and perturbation levels. The Parallel GCNN model consistently achieved the best performance across most corruption types, especially for lower perturbation thresholds. Parallel GCNN with Rotation and Scale Equivariance does not perform so well compared with baseline model under the data corruption. We assess the robustness of the proposed Parallel GCNN models using the maximum invariant perturbation metric [112], which quantifies the largest input perturbations a model can tolerate without altering the model's prediction. To further understand the contribution of each equivariant convolutional module, we perform ablation studies under default settings. The experimental details and visualizations of these studies are provided in Appendix F.

7 Conclusion

In this work, we conducted a systematic investigation into the role of architectural symmetry enforcement in improving adversarial robustness. By incorporating rotation- and scale-equivariant convolutions into standard CNNs, we demonstrated that symmetry-aware models could achieve improved resilience against adversarial attacks without relying on adversarial training or extensive data augmentation. Our theoretical analysis showed that equivariant architectures reduced hypothesis space complexity, regularized gradient behavior, and yielded tighter CLEVER-certified robustness bounds. These models consistently preserved classification margins under group transformations and suppressed gradient sensitivity in directions aligned with the data manifold's symmetry structure. Future work could extend these insights to larger-scale datasets, broader threat models, and more expressive network architectures.

Acknowledgment

This work was supported in part by the National Science Foundation under Award #2346643, OAC-2313191 and by the National Research Platform (NRP) Nautilus HPC cluster [113].

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[3] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, J Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[4] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114, 2019.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[6] Mingli Zhu, Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Junhao Dong, Piotr Koniusz, Junxi Chen, and Yew-Soon Ong. Adversarially robust distillation by reducing the student-teacher variance gap. In *European Conference on Computer Vision*, pages 92–111. Springer, 2025.

[8] Taco S Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2990–2999, 2016.

[9] Daniel E Worrall and Max Welling. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

[10] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] Joao Esteves, Ameesh Makadia, Kostas Daniilidis, and Adrien Gaidon. 3d object classification and retrieval with spherical cnns. In *Advances in Neural Information Processing Systems*, pages 12759–12770, 2018.

[12] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations*, 2018.

[13] Z. Zhou, Y. Zhou, Z. Zhang, L. Lyu, D. Yan, R. Jin, and D. Dou. Flexible, efficient, and stable adversarial attacks on machine unlearning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML'25)*, Vancouver, Canada, July 13–19 2025.

[14] L. Wang, M. Nayyem, A. A. Rakin, K. Santosh, C. Zhang, and Y. Zhou. Explainability-guided defense: Attribution-aware model refinement against adversarial data attacks. In *Proceedings of the 25th IEEE International Conference on Data Mining (ICDM'25)*, Washington, DC, November 12–15 2025.

[15] J. Jia, J. Liu, C. Huo, Y. Shen, Y. Zhou, H. Dai, and D. Dou. Efficient federated learning with timely update dissemination. *Knowledge and Information Systems (KAIS)*, 2025.

[16] J. Liu, B. Ma, Q. Yu, R. Jin, J. Zhou, Y. Zhou, H. Dai, H. Wang, D. Dou, and P. Valduriez. Efficient federated learning with heterogeneous data and adaptive dropout. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 19(8):146:1–146:31, 2025.

[17] B. Palanisamy, L. Liu, Y. Zhou, and Q. Wang. Privacy-preserving publishing of multilevel utility-controlled graph datasets. *ACM Transactions on Internet Technology (TOIT)*, 18(2): 24:1–24:21, 2018.

[18] Y. Zhou, J. Ren, D. Dou, R. Jin, J. Zheng, and K. Lee. Robust meta network embedding against adversarial attacks. In *Proceedings of the 20th IEEE International Conference on Data Mining (ICDM'20)*, pages 1448–1453, Sorrento, Italy, November 17–20 2020.

[19] Z. Zhang, Z. Zhang, Y. Zhou, Y. Shen, R. Jin, and D. Dou. Adversarial attacks on deep graph matching. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, Virtual, December 6–12 2020.

[20] Y. Zhou, Z. Zhang, S. Wu, V. Sheng, X. Han, Z. Zhang, and R. Jin. Robust network alignment via attack signal scaling and adversarial perturbation elimination. In *Proceedings of the 30th Web Conference (WWW'21)*, pages 3884–3895, Virtual Event / Ljubljana, Slovenia, April 19–23 2021.

[21] X. Zhao, Z. Zhang, Z. Zhang, L. Wu, J. Jin, Y. Zhou, R. Jin, D. Dou, and D. Yan. Expressive 1-lipschitz neural networks for robust multiple graph learning against adversarial attacks. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pages 12719–12735, Virtual Event, July 18–24 2021.

[22] J. Ren, Z. Zhang, J. Jin, X. Zhao, S. Wu, Y. Zhou, Y. Shen, T. Che, R. Jin, and D. Dou. Integrated defense for resilient graph matching. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pages 8982–8997, Virtual Event, July 18–24 2021.

[23] Z. Zhang, Z. Zhang, Y. Zhou, L. Wu, S. Wu, X. Han, D. Dou, T. Che, and D. Yan. Adversarial attack against cross-lingual knowledge graph alignment. In *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, pages 5320–5337, Virtual Event / Punta Cana, Dominican Republic, November 7–11 2021.

[24] Z. Zhang, J. Jin, Z. Zhang, Y. Zhou, X. Zhao, J. Ren, J. Liu, L. Wu, R. Jin, and D. Dou. Validating the lottery ticket hypothesis with inertial manifold theory. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pages 30196–30210, Virtual, December 6–14 2021.

[25] G. Zhang, Y. Zhou, S. Wu, Z. Zhang, and D. Dou. Cross-lingual entity alignment with adversarial kernel embedding and adversarial knowledge translation. *CoRR*, abs/2104.07837, 2021.

[26] Y. Zhou, J. Ren, R. Jin, Z. Zhang, J. Zheng, Z. Jiang, D. Yan, and D. Dou. Unsupervised adversarial network alignment with reinforcement learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3):50:1–50:29, 2022.

[27] J. Jin, Z. Zhang, Y. Zhou, and L. Wu. Input-agnostic certified group fairness via gaussian parameter smoothing. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, pages 10340–10361, Baltimore, MD, July 17–23 2022.

[28] Z. Zhang, Y. Zhou, X. Zhao, T. Che, and L. Lyu. Prompt certified machine unlearning with randomized gradient smoothing and quantization. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, New Orleans, LA, November 28–December 9 2022.

[29] J. Jin, J. Ren, Y. Zhou, L. Lv, J. Liu, and D. Dou. Accelerated federated learning with decoupled adaptive optimization. In *Proceedings of the 39th International Conference on Machine Learning (ICML'22)*, pages 10298–10322, Baltimore, MD, July 17–23 2022.

[30] T. Che, Z. Zhang, Y. Zhou, X. Zhao, J. Liu, Z. Jiang, D. Yan, R. Jin, and D. Dou. Federated fingerprint learning with heterogeneous architectures. In *Proceedings of the 22nd IEEE International Conference on Data Mining (ICDM'22)*, pages 31–40, Orlando, FL, November 28–December 1 2022.

[31] J. Liu, J. Huang, Y. Zhou, X. Li, S. Ji, H. Xiong, and D. Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems (KAIS)*, 64(4): 885–917, 2022.

[32] T. Che, Y. Zhou, Z. Zhang, L. Lyu, J. Liu, D. Yan, D. Dou, and J. Huan. Fast federated machine unlearning with nonlinear functional theory. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, pages 4241–4268, Honolulu, HI, July 23–29 2023.

[33] J. Ren, Y. Zhou, J. Jin, L. Lyu, and D. Yan. Dimension-independent certified neural network watermarks via mollifier smoothing. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, pages 28976–29008, Honolulu, HI, July 23–29 2023.

[34] T. Che, J. Liu, Y. Zhou, J. Ren, J. Zhou, V. S. Sheng, H. Dai, and D. Dou. Federated learning of large language models with parameter-efficient prompt tuning and adaptive optimization. In *Proceedings of the 28th Conference on Empirical Methods in Natural Language Processing (EMNLP'23)*, Singapore, December 6–10 2023.

[35] J. Liu, J. Jia, B. Ma, C. Zhou, J. Zhou, Y. Zhou, H. Dai, and D. Dou. Multi-job intelligent scheduling with cross-device federated learning. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 34(2):535–551, 2023.

[36] J. Liu, J. Jia, T. Che, C. Huo, J. Ren, Y. Zhou, H. Dai, and D. Dou. FedASMU: Efficient asynchronous federated learning with dynamic staleness-aware model update. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI'24)*, pages 13900–13908, Vancouver, Canada, February 20–27 2024.

[37] J. Liu, T. Che, Y. Zhou, R. Jin, H. Dai, D. Dou, and P. Valduriez. AEDFL: Efficient asynchronous decentralized federated learning with heterogeneous devices. In *Proceedings of the 24th SIAM International Conference on Data Mining (SDM'24)*, pages 833–841, Houston, TX, April 18–20 2024.

[38] Y. Zhou, Z. Zhang, Z. Zhang, L. Lyu, and W.-S. Ku. Effective federated graph matching. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Vienna, Austria, July 21–27 2024.

[39] Y. Xiao, Z. Zhang, Y. Fang, D. Yan, Y. Zhou, W. Ku, and B. Hui. Advancing certified robustness of explanation via gradient quantization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM'24)*, pages 2596–2606, Boise, ID, October 21–25 2024.

[40] J. Liu, J. Ren, R. Jin, Z. Zhang, Y. Zhou, P. Valduriez, and D. Dou. Fisher information-based efficient curriculum federated learning with large language models. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 10497–10523, Miami, FL, November 12–16 2024.

[41] J. Liu, J. Jia, H. Zhang, Y. Yun, L. Wang, Y. Zhou, H. Dai, and D. Dou. Efficient federated learning using dynamic update and adaptive pruning with momentum on shared server data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 15(6):122:1–122:28, 2024.

[42] Y. Zhou, H. Cheng, and J. X. Yu. Clustering large attributed graphs: An efficient incremental approach. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM'10)*, pages 689–698, Sydney, Australia, December 14–17 2010.

[43] Y. Zhou, H. Cheng, and J. X. Yu. Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment (PVLDB)*, 2(1):718–729, 2009.

[44] H. Cheng, Y. Zhou, and J. X. Yu. Clustering large attributed graphs: A balance between structural and attribute similarities. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(2):1–33, 2011.

[45] Y. Zhou and L. Liu. Clustering analysis in large graphs with rich attributes. In D. E. Holmes and L. C. Jain, editors, *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*. Springer, 2012.

[46] H. Cheng, Y. Zhou, X. Huang, and J. X. Yu. Clustering large attributed information networks: An efficient incremental computing approach. *Data Mining and Knowledge Discovery (DMKD)*, 25(3):450–477, 2012.

[47] Longwei Wang and Qilian Liang. Representation learning and nature encoded fusion for heterogeneous sensor networks. *IEEE Access*, 7:39227–39235, 2019.

[48] Longwei Wang, Wen Chen, and Jun Li. Congestion aware dynamic user association in heterogeneous cellular network: A stochastic decision approach. In *2014 IEEE International Conference on Communications (ICC)*, pages 2636–2640. IEEE, 2014.

[49] Longwei Wang, Aashish Ghimire, K Santosh, Zheng Zhang, and Xueqian Li. Enhanced robustness by symmetry enforcement. *IEEE CAI*, 2024.

[50] Y. Zhou and L. Liu. Social influence based clustering of heterogeneous information networks. In *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'13)*, pages 338–346, Chicago, IL, August 11–14 2013.

[51] Z. Su, L. Liu, M. Li, X. Fan, and Y. Zhou. ServiceTrust: Trust management in service provision networks. In *Proceedings of the 10th IEEE International Conference on Services Computing (SCC'13)*, pages 272–279, Santa Clara, CA, June 27–July 2 2013.

[52] Q. Zhang, L. Liu, Y. Ren, K. Lee, Y. Tang, X. Zhao, and Y. Zhou. Residency aware inter-VM communication in virtualized cloud: Performance measurement and analysis. In *Proceedings of the 2013 IEEE International Conference on Cloud Computing (CLOUD'13)*, pages 204–211, Santa Clara, CA, June 27–July 2 2013.

[53] Y. Zhou and L. Liu. Activity-edge centric multi-label classification for mining heterogeneous information networks. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'14)*, pages 1276–1285, New York, NY, August 24–27 2014.

[54] Z. Su, L. Liu, M. Li, X. Fan, and Y. Zhou. Reliable and resilient trust management in distributed service provision networks. *ACM Transactions on the Web (TWEB)*, 9(3):1–37, 2015.

[55] Y. Zhou and L. Liu. Social influence based clustering and optimization over heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1): 1–53, 2015.

[56] Y. Zhou, L. Liu, and D. Buttler. Integrating vertex-centric clustering with edge-centric clustering for meta path graph analysis. In *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'15)*, pages 1563–1572, Sydney, Australia, August 10–13 2015.

[57] X. Bao, L. Liu, N. Xiao, Y. Zhou, and Q. Zhang. Policy-driven autonomic configuration management for nosql. In *Proceedings of the 2015 IEEE International Conference on Cloud Computing (CLOUD'15)*, pages 245–252, New York, NY, June 27–July 2 2015.

[58] Y. Zhou, L. Liu, K. Lee, C. Pu, and Q. Zhang. Fast iterative graph computation with resource aware graph parallel abstractions. In *Proceedings of the 24th ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC'15)*, pages 179–190, Portland, OR, June 15–19 2015.

[59] Y. Zhou, L. Liu, S. Seshadri, and L. Chiu. Analyzing enterprise storage workloads with graph modeling and clustering. *IEEE Journal on Selected Areas in Communications (JSAC)*, 34(3): 551–574, 2016.

[60] Y. Zhou. *Innovative Mining, Processing, and Application of Big Graphs*. PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2017.

[61] Y. Zhou, S. Wu, C. Jiang, Z. Zhang, D. Dou, R. Jin, and P. Wang. Density-adaptive local edge representation learning with generative adversarial network multi-label edge classification. In *Proceedings of the 18th IEEE International Conference on Data Mining (ICDM'18)*, pages 1464–1469, Singapore, November 17–20 2018.

[62] Y. Zhou, A. Amimeur, C. Jiang, D. Dou, R. Jin, and P. Wang. Density-aware local siamese autoencoder network embedding with autoencoder graph clustering. In *Proceedings of the 2018 IEEE International Conference on Big Data (BigData'18)*, pages 1162–1167, Seattle, WA, December 10–13 2018.

[63] Longwei Wang, Chengfei Wang, Yupeng Li, and Rui Wang. Explaining the behavior of neuron activations in deep neural networks. *Ad Hoc Networks*, 111:102346, 2021.

[64] Longwei Wang, Chengfei Wang, Yupeng Li, and Rui Wang. Improving robustness of deep neural networks via large-difference transformation. *Neurocomputing*, 450:411–419, 2021.

[65] Kenan Xiao, Longwei Wang, Ashish Gupta, and Xiao Qin. Looking beyond content: Modeling and detection of fake news from a social context perspective. In *HICSS*, pages 1–10, 2022.

[66] J. Ren, Y. Zhou, R. Jin, Z. Zhang, D. Dou, and P. Wang. Dual adversarial learning based network alignment. In *Proceedings of the 19th IEEE International Conference on Data Mining (ICDM'19)*, pages 1288–1293, Beijing, China, November 8–11 2019.

[67] Y. Zhou, L. Liu, Q. Zhang, and B. Palanisamy. Enhancing collaborative filtering with multi-label classification. In *Proceedings of the 2019 International Conference on Computational Data and Social Networks (CSoNet'19)*, pages 323–338, Ho Chi Minh City, Vietnam, November 18–20 2019.

[68] Y. Zhou, C. Jiang, Z. Zhang, D. Dou, R. Jin, and P. Wang. Integrating local vertex/edge embedding via deep matrix fusion and siamese multi-label classification. In *Proceedings of the 2019 IEEE International Conference on Big Data (BigData'19)*, pages 1018–1027, Los Angeles, CA, December 9–12 2019.

[69] Y. Zhou, J. Ren, S. Wu, D. Dou, R. Jin, Z. Zhang, and P. Wang. Semi-supervised classification-based local vertex ranking via dual generative adversarial nets. In *Proceedings of the 2019 IEEE International Conference on Big Data (BigData'19)*, pages 1267–1273, Los Angeles, CA, December 9–12 2019.

[70] Y. Zhou and L. Liu. Approximate deep network embedding for mining large-scale graphs. In *Proceedings of the 2019 IEEE International Conference on Cognitive Machine Intelligence (CogMI'19)*, pages 53–60, Los Angeles, CA, December 12–14 2019.

[71] S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, pages 5811–5820, Online, July 5–10 2020.

[72] S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu. TopicKA: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)*, pages 3766–3772, Online, January 7–15 2021.

[73] Y. Zhou, J. Ren, R. Jin, Z. Zhang, D. Dou, and D. Yan. Unsupervised multiple network alignment with multinominal GAN and variational inference. In *Proceedings of the 2020 IEEE International Conference on Big Data (BigData'20)*, pages 868–877, Atlanta, GA, December 10–13 2020.

[74] Longwei Wang, Peijie Chen, Chengfei Wang, and Rui Wang. Layer-wise entropy analysis and visualization of neurons activation. In *International Conference on Communications and Networking in China*, pages 29–36. Springer International Publishing Cham, 2019.

[75] Longwei Wang, Xueqian Li, and Zheng Zhang. Dense cross-connected ensemble convolutional neural networks for enhanced model robustness. *arXiv preprint arXiv:2412.07022*, 2024.

[76] Navid Nayyem, Abdullah Rakin, and Longwei Wang. Bridging interpretability and robustness using lime-guided model refinement. *arXiv preprint arXiv:2412.18952*, 2024.

[77] Longwei Wang, Navid Nayyem, and Abdullah Rakin. Enhancing adversarial robustness of deep neural networks through supervised contrastive learning. *arXiv preprint arXiv:2412.19747*, 2024.

[78] Longwei Wang, Ifrat Ikhtear Uddin, Xiao Qin, Yang Zhou, and KC Santosh. Explainability-driven defense: Grad-cam-guided model refinement against adversarial threats. In *Proceedings of the AAAI Symposium Series*, volume 6, pages 49–57, 2025.

[79] Robin Narsingh Ranabhat, Longwei Wang, Xiao Qin, Yang Zhou, and KC Santosh. Multi-scale unrectified push-pull with channel attention for enhanced corruption robustness. In *Proceedings of the AAAI Symposium Series*, volume 6, pages 34–41, 2025.

[80] Ifrat Ikhtear Uddin, Longwei Wang, and KC Santosh. Expert-guided explainable few-shot learning for medical image diagnosis. *arXiv preprint arXiv:2509.08007*, 2025.

[81] Y. Zhou, L. Liu, K. Lee, B. Palanisamy, and Q. Zhang. Improving collaborative filtering with social influence over heterogeneous information networks. *ACM Transactions on Internet Technology (TOIT)*, 20(4):36:1–36:29, 2020.

[82] R. Jin, D. Li, J. Gao, Z. Liu, L. Chen, and Y. Zhou. Towards a better understanding of linear models for recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'21)*, pages 776–785, Virtual Event, Singapore, August 14–18 2021.

[83] Robin Narsingh Ranabhat, Longwei Wang, Amit Kumar Patel, et al. Promoting shape bias in cnns: Frequency-based and contrastive regularization for corruption robustness. *arXiv preprint arXiv:2509.11355*, 2025.

[84] S. Wu, M. Wang, D. Zhang, Y. Zhou, Y. Li, and Z. Wu. Knowledge-aware dialogue generation via hierarchical infobox accessing and infobox-dialogue interaction graph network. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, pages 3964–3970, Virtual Event / Montreal, Canada, August 19–27 2021.

[85] S. Wu, Y. Li, M. Wang, D. Zhang, Y. Zhou, and Z. Wu. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, pages 2286–2300, Virtual Event / Punta Cana, Dominican Republic, November 7–11 2021.

[86] Guimu Guo, Da Yan, Yang Zhou, et al. Maximal directed quasi-clique mining. In *Proceedings of the 38th IEEE International Conference on Data Engineering (ICDE'22)*, pages 1900–1913, Kuala Lumpur, Malaysia, May 9–12 2022.

[87] Zhiqiang Chen, Yang Chen, Xiaolong Zou, and Shan Yu. Continuous rotation group equivariant network inspired by neural population coding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11462–11470, 2024.

[88] Hayder Elesedy. Symmetry and generalisation in machine learning. *arXiv preprint arXiv:2501.03858*, 2025.

[89] Sharvaree Vadgama, Mohammad Mohaiminul Islam, Domas Buracus, Christian Shewmake, and Erik Bekkers. On the utility of equivariance and symmetry breaking in deep learning architectures on point clouds. *arXiv preprint arXiv:2501.01999*, 2025.

[90] Daniel E Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. *arXiv preprint arXiv:1905.11696*, 2019.

[91] Maurice Weiler and Giacomo Cesa. General e(2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pages 14334–14345, 2019.

[92] Ramzan Basheer and Deepak Mishra. Current symmetry group equivariant convolution frameworks for representation learning. *arXiv preprint arXiv:2409.07327*, 2024.

[93] Maksim Zhdanov, Nabil Iqbal, Erik J Bekkers, and Patrick Forré. Ads-gnn-a conformally equivariant graph neural network. In *ICLR 2025 Workshop on Machine Learning Multiscale Processes*.

[94] Kristof T Schütt, Pieter-Jan Kindermans, Huziel E Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pages 992–1002, 2017.

[95] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[96] Jon Vadillo, Roberto Santana, and Jose A Lozano. Adversarial attacks in explainable machine learning: A survey of threats against models and humans. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1):e1567, 2025.

[97] Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pages 461–478. Springer, 2025.

[98] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019.

[99] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

[100] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, 2024.

[101] Sheng Y Peng, Pin-Yu Chen, Matthew Hull, and Duen H Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. *Advances in Neural Information Processing Systems*, 37:95692–95715, 2024.

[102] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pages 274–283, 2018.

[103] Desheng Zheng, Wuping Ke, Xiaoyu Li, Yaoxin Duan, Guangqiang Yin, and Fan Min. Enhancing the transferability of adversarial attacks via multi-feature attention. *IEEE Transactions on Information Forensics and Security*, 2025.

[104] Xu Zhang, Bo Peng, Jianjun Lei, Chao Xue, Yuxuan Yao, and Qingming Huang. Adversarially robust object detection via deviation calibration and content preservation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[105] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[106] Ling Wang, Chuan Wang, Yuan Li, and Rui Wang. Improving robustness of deep neural networks via large-difference transformation. *Neurocomputing*, 450:411–419, 2021.

[107] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

[108] Fabio Anselmi, Georgios Evangelopoulos, Lorenzo Rosasco, and Tomaso Poggio. Symmetry-adapted representation learning. *Pattern Recognition*, 86:201–208, 2019.

[109] Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On invariance and selectivity in representation learning. *Information and Inference: A Journal of the IMA*, 5(2):134–158, 2016.

[110] Fabio Anselmi, Joel Z Leibo, Lorenzo Rosasco, Jim Mutch, Andrea Tacchetti, and Tomaso Poggio. Unsupervised learning of invariant representations. *Theoretical Computer Science*, 633:112–121, 2016.

[111] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[112] Satoshi Hara, Kouichi Ikeno, Tasuku Soma, and Takanori Maehara. Maximally invariant data perturbation as explanation. *arXiv preprint arXiv:1806.07004*, 2018.

[113] Derek Weitzel, Ashton Graves, Sam Albin, Huijun Zhu, Frank Wuerthwein, Mahidhar Tatineni, Dmitry Mishin, Elham Khoda, Mohammad Sada, Larry Smarr, et al. The national research platform: Stretched, multi-tenant, scientific kubernetes cluster. In *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, pages 1–5, 2025.

A Technical Appendices and Supplementary Material

A.1 Proof of Theorem 1

Proof. We begin by computing the gradient of the margin function $g_{c,j}$ at the transformed input $g \cdot \mathbf{x}$. Given the equivariance property:

$$f(g \cdot \mathbf{x}) = \rho(g) f(\mathbf{x}),$$

we write the individual logits as:

$$f_i(g \cdot \mathbf{x}) = [\rho(g) f(\mathbf{x})]_i = \sum_{m=1}^{k} \rho_{im}(g) f_m(\mathbf{x}).$$

Assuming $\rho(g)$ is diagonal (as in classification tasks where each logit transforms independently), the above reduces to:

$$f_i(g \cdot \mathbf{x}) = \rho_{ii}(g) f_i(\mathbf{x}).$$

Thus, the margin function transforms as:

$$g_{c,j}(g \cdot \mathbf{x}) = f_c(g \cdot \mathbf{x}) - f_j(g \cdot \mathbf{x}) = \rho_{cc}(g) f_c(\mathbf{x}) - \rho_{jj}(g) f_j(\mathbf{x}).$$

Now compute the gradient using the chain rule. Since $g \cdot \mathbf{x}$ is a diffeomorphism and f is differentiable, we have:

$$\nabla g_{c,j}(g \cdot \mathbf{x}) = \nabla_{\mathbf{z}} g_{c,j}(\mathbf{z})\big|_{\mathbf{z}=g\cdot\mathbf{x}} = \nabla g_{c,j}(g \cdot \mathbf{x}) = J_{g_{c,j}}(g \cdot \mathbf{x}) = \nabla \left(f_c(g \cdot \mathbf{x}) - f_j(g \cdot \mathbf{x}) \right).$$

Applying the chain rule to each term:

$$\nabla f_c(g \cdot \mathbf{x}) = \nabla \left(\rho_{cc}(g) f_c(\mathbf{x}) \right) = \rho_{cc}(g) \cdot \nabla f_c(\mathbf{x}) \cdot Dg^{-1},$$
$$\nabla f_j(g \cdot \mathbf{x}) = \rho_{jj}(g) \cdot \nabla f_j(\mathbf{x}) \cdot Dg^{-1}.$$

Hence:

$$\begin{aligned} \nabla g_{c,j}(g \cdot \mathbf{x}) &= \nabla f_c(g \cdot \mathbf{x}) - \nabla f_j(g \cdot \mathbf{x}) \\ &= \rho_{cc}(g) \nabla f_c(\mathbf{x}) Dg^{-1} - \rho_{jj}(g) \nabla f_j(\mathbf{x}) Dg^{-1} \\ &= \left(\rho_{cc}(g) \nabla f_c(\mathbf{x}) - \rho_{jj}(g) \nabla f_j(\mathbf{x}) \right) Dg^{-1}. \end{aligned}$$

We now compute the ℓ_q-norm:

$$\|\nabla g_{c,j}(g \cdot \mathbf{x})\|_q = \left\| \left(\rho_{cc}(g) \nabla f_c(\mathbf{x}) - \rho_{jj}(g) \nabla f_j(\mathbf{x}) \right) Dg^{-1} \right\|_q.$$

Assume that $\rho_{cc}(g), \rho_{jj}(g) \in \{+1, -1\}$ and that Dg^{-1} is orthogonal:

$$\|Dg^{-1}\mathbf{v}\|_q = \|\mathbf{v}\|_q, \quad \forall \mathbf{v} \in \mathbb{R}^d.$$

Then:

$$\begin{aligned} \|\nabla g_{c,j}(g \cdot \mathbf{x})\|_q &= \left\| \left(\rho_{cc}(g) \nabla f_c(\mathbf{x}) - \rho_{jj}(g) \nabla f_j(\mathbf{x}) \right) Dg^{-1} \right\|_q \\ &= \left\| \rho_{cc}(g) \nabla f_c(\mathbf{x}) - \rho_{jj}(g) \nabla f_j(\mathbf{x}) \right\|_q \\ &= \|\nabla g_{c,j}(\mathbf{x})\|_q. \end{aligned}$$

Therefore, the norm of the gradient of the margin function is invariant under the group action:

$$\|\nabla g_{c,j}(g \cdot \mathbf{x})\|_q = \|\nabla g_{c,j}(\mathbf{x})\|_q, \quad \forall g \in G.$$

Since this holds for every $g \in G$, the Lipschitz constant of $g_{c,j}$ over the group orbit $[\mathbf{x}]_G$ satisfies:

$$L_q^{(j)} = \sup_{\mathbf{x}' \in B_p([\mathbf{x}]_G, r)} \|\nabla g_{c,j}(\mathbf{x}')\|_q = \|\nabla g_{c,j}(\mathbf{x})\|_q.$$

This completes the proof. □

A.2 Proof of Theorem 2

Proof. The key idea is that group equivariance induces structured regularity over the orbit of the input space. For any $g \in G$, by applying the chain rule to the equivariant condition $f(g \cdot \mathbf{x}) = \rho(g)f(\mathbf{x})$, we obtain the Jacobian transformation law:

$$J_f(g \cdot \mathbf{x}) = \rho(g)J_f(\mathbf{x})Dg^{-1}, \tag{2}$$

where $Dg^{-1} \in \mathbb{R}^{d\times d}$ denotes the Jacobian of the inverse group action $g^{-1} \cdot \mathbf{x}$. Because both $\rho(g)$ and Dg^{-1} are orthogonal, the Jacobian norm is preserved:

$$\|J_f(g \cdot \mathbf{x})\|_2 = \|J_f(\mathbf{x})\|_2 . \tag{3}$$

Now define the orbit-averaged gradient for each class logit $f_j(\mathbf{x})$ as:

$$\overline{\nabla f_j}(\mathbf{x}) := \frac{1}{|G|}\sum_{g\in G} \nabla f_j(g \cdot \mathbf{x}). \tag{4}$$

Using Eq. (2), we get:

$$\overline{\nabla f_j}(\mathbf{x}) = \frac{1}{|G|}\sum_{g\in G} \rho(g)\nabla f_j(\mathbf{x})Dg^{-1}. \tag{5}$$

This is a linear averaging operator acting on $\nabla f_j(\mathbf{x})$, which has the effect of projecting the gradient onto the invariant subspace under the group. High-frequency components in directions orthogonal to $T_{\mathbf{x}}([\mathbf{x}]_G)$ tend to cancel due to the group symmetries, while components aligned with the orbit are reinforced.

We now analyze the gradient difference along perturbations in each direction.

Case 1: Perturbation $\boldsymbol{\delta}_G \in T_{\mathbf{x}}([\mathbf{x}]_G)$:

Let $\mathbf{x}' = \mathbf{x}+\boldsymbol{\delta}_G$. Since $\boldsymbol{\delta}_G$ lies tangent to the orbit, we have $\mathbf{x}' \in [\mathbf{x}]_G$, and hence by the equivariance property and Eq. (3),

$$\nabla f(\mathbf{x}') \approx \nabla f(\mathbf{x}),$$

and thus,

$$\|\nabla f(\mathbf{x} + \boldsymbol{\delta}_G) - \nabla f(\mathbf{x})\|_2 \approx 0.$$

—

Case 2: Perturbation $\boldsymbol{\delta}_\perp \perp T_{\mathbf{x}}([\mathbf{x}]_G)$:

Let $\mathbf{x}'' = \mathbf{x} + \boldsymbol{\delta}_\perp$. Because this point lies **off the symmetry manifold**, the equivariant structure offers no constraint on the variation of the gradient. Consequently,

$$\|\nabla f(\mathbf{x} + \boldsymbol{\delta}_\perp) - \nabla f(\mathbf{x})\|_2$$

is not bounded by symmetry and may vary substantially—especially in directions where ∇f contains high-frequency components not captured by the group action.

Moreover, the difference between smoothed and unsmoothed gradients satisfies:

$$\left\|\overline{\nabla f_j}(\mathbf{x}) - \nabla f_j(\mathbf{x})\right\|_2 \ll \|\nabla f_j(\mathbf{x} + \boldsymbol{\delta}_\perp) - \nabla f_j(\mathbf{x})\|_2 . \tag{6}$$

This shows that the orbit-averaged gradient stabilizes variation, while off-orbit perturbations significantly alter the local gradient field.

Combining Eqs. (A.2) and (6), we conclude:

$$\|\nabla f(\mathbf{x} + \boldsymbol{\delta}_\perp) - \nabla f(\mathbf{x})\|_2 \gg \|\nabla f(\mathbf{x} + \boldsymbol{\delta}_G) - \nabla f(\mathbf{x})\|_2 ,$$

which proves the theorem. □

B Limitations

Despite its theoretical and empirical strengths, this work has several limitations. First, the theoretical framework focuses solely on local Lipschitz bounds (e.g., CLEVER), which may not capture broader robustness phenomena such as margin distributions or adversarial risk. Second, empirical evaluations are limited to small-scale datasets and common ℓ_p-norm attacks, leaving open questions about scalability and robustness to more diverse threat models. Finally, the integration of equivariant layers is confined to early stages of the network, and the computational trade-offs of group convolutions remain unquantified.

C Equivariance Enhanced Architectural Designs

C.1 Group Equivariant Convolutions

Group Equivariant Convolutional Networks (G-CNNs) extend the conventional convolutional framework by embedding symmetry priors directly into the architecture [8]. These models are designed to maintain equivariance under transformations defined by a group G, such as rotations, translations, or scalings.

To implement such equivariance in convolutional neural networks, the standard convolution operation is replaced with a *group convolution* that aggregates information over the transformed versions of both input and filters. We now instantiate this framework for two practically important cases: rotation and scale transformations.

C.1.1 Rotation-Equivariant Convolutions

To embed rotational symmetry into the network, we consider a discrete rotation group $G = P4$, which consists of four planar rotations: $0°, 90°, 180°$, and $270°$. In this setting, group convolution is defined as:

$$[f * g](u) = \sum_{v \in G} f(v^{-1}u)g(v), \tag{7}$$

where f is the input feature map, $g(v)$ is the group-transformed filter corresponding to transformation $v \in G$, and u denotes the spatial coordinate. This formulation guarantees that the resulting feature map transforms predictably under group actions:

$$[f * g](g \cdot u) = \rho(g)[f * g](u), \quad \forall g \in G, \tag{8}$$

thereby satisfying the equivariance condition in Equation (1). By construction, these layers promote feature consistency across rotated versions of the input, enhancing robustness and generalization.

C.1.2 Scale-Equivariant Convolutions

Scale transformations are another common form of variability in visual data, especially when objects appear at different sizes or distances. To build scale-equivariant architectures, we define a discrete scale group $G_s = \{\alpha_1, \alpha_2, \ldots, \alpha_k\} \subset \mathbb{R}^+$, which acts on the input space via isotropic resizing:

$$\alpha \cdot x := \text{Resize}(x, \alpha), \quad \alpha \in G_s, \tag{9}$$

where $\text{Resize}(x, \alpha)$ uniformly scales the spatial dimensions of the input x by a factor α, typically using bicubic interpolation. Each scaled input is then processed independently using a shared convolutional function f_{scale}, resulting in multiple feature maps:

$$F_{\text{scale},i} = f_{\text{scale}}(\alpha_i \cdot x), \quad i = 1, \ldots, k. \tag{10}$$

These outputs are then resized back to a common resolution and combined via an aggregation operation, such as channel-wise concatenation or averaging:

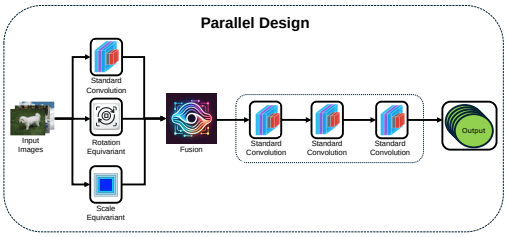

Figure 3: Parallel design of the CNN architecture. Input is processed through standard, rotation-equivariant, and scale-equivariant branches independently. The resulting features are subsequently fused to form a unified representation.

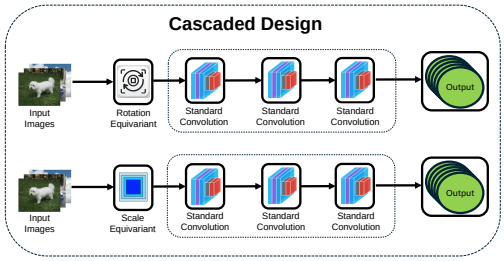

Figure 4: Cascaded design of the CNN architecture. Convolutional layers are applied sequentially: standard, then rotation-equivariant, followed by scale-equivariant.

$$F_{\text{scale}}(x) = \text{Aggregate}(F_{\text{scale},1}, \ldots, F_{\text{scale},k}). \tag{11}$$

The resulting representation is equivariant to the scale group G_s, satisfying:

$$F_{\text{scale}}(\beta \cdot x) = \rho(\beta) F_{\text{scale}}(x), \quad \forall \beta \in G_s, \tag{12}$$

where $\rho(\beta)$ is the corresponding transformation (e.g., channel permutation or spatial rescaling) applied to the feature space. This construction allows the network to maintain consistent semantic representations across a range of object sizes, improving its ability to generalize under scale variation.

Both rotation- and scale-equivariant convolutions are special cases of the general group-equivariant convolutional framework. They embed structured inductive biases into the network that enable efficient learning and enhanced generalization.

C.2 Equivariance Enforced Architectural Designs

We explore two architectural strategies that combine standard, rotation-equivariant, and scale-equivariant convolutions for robust feature extraction: the *parallel* design and the *cascaded* design. Each strategy targets a different trade-off between representation diversity and processing efficiency.

C.3 Parallel Design

In the parallel design, the input is processed simultaneously through three branches: a standard convolutional branch for translational equivariant features, a rotation-equivariant branch that encodes orientation equivariance via group convolutions, and a scale-equivariant branch that captures multi-scale features.

Let $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ denote the input image, where C, H, and W are the number of channels and the spatial height and width, respectively. The outputs from each branch are given by:

$$\mathbf{F}_{\text{standard}} = f_{\text{standard}}(\mathbf{x}), \quad \mathbf{F}_{\text{rot}} = f_{\text{rot}}(\mathbf{x}), \quad \mathbf{F}_{\text{scale}} = f_{\text{scale}}(\mathbf{x}), \tag{13}$$

where f_{standard}, f_{rot}, and f_{scale} are the respective processing functions. The final representation is obtained by fusing the branch outputs:

$$\mathbf{F}_{\text{fused}} = g(\mathbf{F}_{\text{standard}}, \mathbf{F}_{\text{rot}}, \mathbf{F}_{\text{scale}}), \tag{14}$$

where $g(\cdot)$ denotes the fusion strategy. Figure 3 illustrates this design, emphasizing the independence of the branches and their contribution to robust feature diversity.

C.4 Cascaded Design

In contrast to the parallel strategy, the cascaded design applies equivariant operations sequentially. The input first passes through a standard convolutional layer, followed by rotation-equivariant processing,

Table 2: Adversarial robustness of fully equivariant architectures on CIFAR-10 (%). Deeper models benefit from compounding gradient regularization when equivariance is enforced end-to-end.

Architecture	$\varepsilon = 0.01$		$\varepsilon = 0.03$		$\varepsilon = 0.05$		$\varepsilon = 0.10$	
	FGSM	PGD	FGSM	PGD	FGSM	PGD	FGSM	PGD
4-Layer Equivariant	65.65	52.20	53.78	23.30	47.08	15.85	37.95	7.01
10-Layer Equivariant	**73.01**	**64.96**	**67.09**	**52.37**	**60.23**	**37.80**	**44.93**	**12.46**

Table 3: Adversarial robustness of fully equivariant architectures on CIFAR-100 (%). Consistent improvements across depth validate that orbit-invariant gradient regularization scales effectively.

Architecture	$\varepsilon = 0.01$		$\varepsilon = 0.03$		$\varepsilon = 0.05$		$\varepsilon = 0.10$	
	FGSM	PGD	FGSM	PGD	FGSM	PGD	FGSM	PGD
4-Layer Equivariant	38.40	21.59	26.63	8.92	22.06	5.66	15.96	2.39
10-Layer Equivariant	**50.60**	**36.29**	**42.02**	**21.34**	**36.09**	**12.01**	**24.68**	**4.08**

and finally scale-equivariant processing:

$$\mathbf{F}_{\text{rot}} = f_{\text{rot}}(f_{\text{standard}}(\mathbf{x})), \quad \mathbf{F}_{\text{scale}} = f_{\text{scale}}(\mathbf{F}_{\text{rot}}), \tag{15}$$

resulting in the final representation:

$$\mathbf{F}_{\text{fused}} = g(\mathbf{F}_{\text{scale}}). \tag{16}$$

This design simplifies the model by reusing intermediate representations. However, it may also introduce feature redundancy or suppress early-layer diversity, as subsequent operations act on already transformed representations. Figure 4 depicts the cascaded architecture, highlighting the flow of information through progressively specialized layers.

D Fully Equivariant Architectures

To validate that our theoretical predictions hold for end-to-end equivariant models, we evaluate architectures where *all* convolutional layers enforce rotation equivariance under the P4 group (4 discrete rotations). These models are constructed by sequentially stacking rotation-equivariant blocks, with each block comprising an equivariant convolution, group-aware batch normalization, ReLU activation, and max pooling.

Tables 2 and 3 demonstrate that fully equivariant models achieve substantial adversarial robustness without adversarial training. The 10-layer architecture consistently outperforms the 4-layer variant across all perturbation levels, with improvements of 7-15% in FGSM accuracy and 10-30% in PGD accuracy at moderate perturbations. This depth-dependent improvement validates our theoretical prediction: when symmetry constraints are enforced throughout the network, orbit-invariant gradient regularization compounds beneficially across layers. This contrasts sharply with standard CNNs, where increased depth often amplifies adversarial vulnerability. The results confirm that equivariance must be treated as a network-wide architectural principle rather than localized feature extraction, demonstrating that certified robustness guarantees translate into measurable improvements under practical gradient-based attacks.

E Comparison with Adversarial Training

To contextualize the robustness improvements from architectural symmetry enforcement, we compare our approach against standard adversarial training. Table 4 presents results on CIFAR-10, where a standard CNN trained with PGD-based adversarial training is compared against our 10-layer fully equivariant G-CNN trained without any adversarial examples.

Our equivariant model achieves superior FGSM robustness and maintains competitive PGD robustness despite not being trained on adversarial examples. This demonstrates that symmetry-aware architectures provide intrinsic robustness through geometric constraints alone, offering a complementary

Table 4: Comparison with Adversarial Training on CIFAR-10.

Model	$\varepsilon = 0.01$		$\varepsilon = 0.02$		$\varepsilon = 0.03$		$\varepsilon = 0.04$		$\varepsilon = 0.05$	
	FGSM	PGD	FGSM	PGD	FGSM	PGD	FGSM	PGD	FGSM	PGD
CNN + AT	74.5	67.0	70.2	60.4	66.1	54.0	61.7	48.3	57.3	42.1
G-CNN (no AT)	73.01	64.96	70.16	58.87	**67.09**	52.37	**63.77**	45.52	**60.23**	37.80

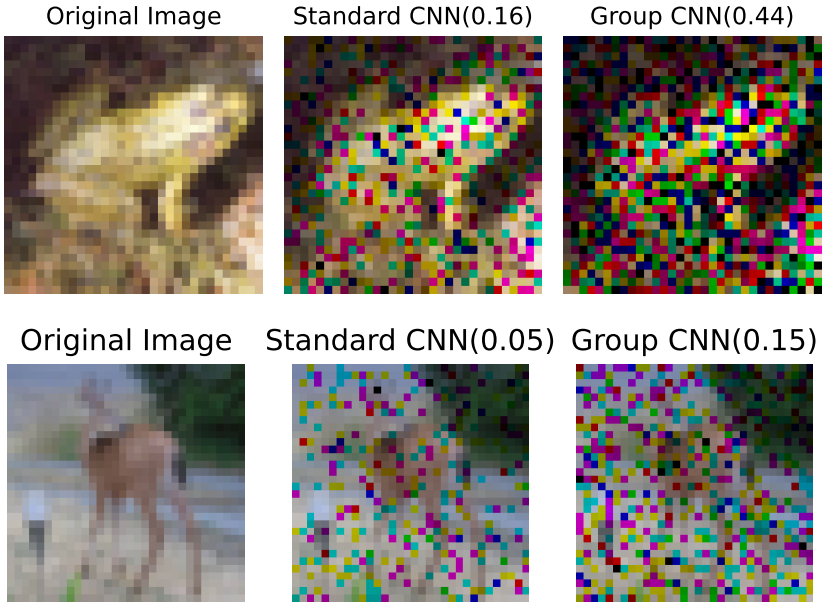

Figure 5: Visualization of perturbation tolerance for the Baseline CNN and the Parallel GCNN. The top row shows the original image (left), followed by the maximum perturbation ($\epsilon = 0.16$) for the Baseline CNN (middle), and the maximum perturbation ($\epsilon = 0.44$) for the Parallel GCNN (right). Similarly, the bottom row shows the original image (left), with the maximum perturbation ($\epsilon = 0.05$) for the Baseline CNN (middle), and the maximum perturbation ($\epsilon = 0.15$) for the Parallel GCNN (right).

and potentially more efficient alternative to data augmentation-based defenses. The slight advantage in FGSM accuracy and competitive PGD performance suggest that equivariant priors effectively regularize the decision boundary without the computational overhead of adversarial training.

F Perturbation Tolerance Visualization and Ablation Study

F.1 Visualization

We evaluated the robustness of the Parallel GCNN models using the metric of *maximum invariant perturbation* [112]. This metric quantifies the highest level of perturbations that a model can withstand without significant degradation in performance. The visualization in the figure compares the perturbation levels tolerated by the baseline CNN and the Parallel GCNN . As shown in the figure 5, the Parallel GCNN model demonstrates a substantially higher tolerance for perturbations compared to the baseline CNN. For example, the baseline CNN can tolerate a perturbation rate of $\epsilon = 0.05$, whereas the Parallel GCNN maintains stable predictions even at higher perturbation levels such as $\epsilon = 0.15$.

F.2 Ablation Study

To evaluate the influence of equivariance convolution modules on the proposed model, we conducted ablation studies in default settings. The experiments compare the performance of four

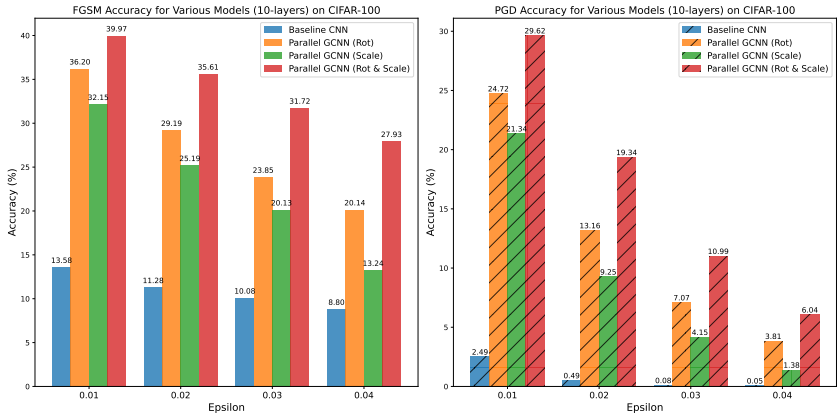

Figure 6: Ablation study results on CIFAR-100 comparing the adversarial robustness of four models under FGSM (left) and PGD (right) attacks.

models—Baseline CNN, Parallel GCNN with Rotation-Equivariant Branch (Rot), Parallel GCNN with Scale-Equivariant Branch (Scale), and Parallel GCNN with Rotation- and Scale-Equivariant Branch (Rot & Scale)—on CIFAR-100 under FGSM and PGD attacks across varying perturbation levels. The Parallel GCNN with Rotation- and Scale-Equivariant Branch consistently outperformed other models, achieving the highest robustness across all settings. Rotation-equivariant convolutions demonstrated better performance than scale-equivariant convolutions individually, particularly under higher perturbation magnitudes. The Baseline CNN lacked adversarial robustness, showing the superiority of symmetry-aware architectural designs.

Under the more challenging PGD attacks, the Parallel GCNN with Rotation- and Scale-Equivariant Branch continued to outperform other models. At $\epsilon = 0.01$, it achieved 29.62% accuracy, compared to 24.72% for the Rotation-Equivariant GCNN, 21.34% for the Scale-Equivariant GCNN, and only 2.49% for the Baseline CNN. Even at $\epsilon = 0.04$, the combined model retained 6.04% accuracy, while the Rotation-Equivariant GCNN and Scale-Equivariant GCNN dropped to 3.81% and 1.38%, respectively. The Baseline CNN showed minimal robustness at higher perturbations, with accuracy falling below 1%.

Our experiments reveal several key insights for the integration of the equivariance layer in the CNN model. The parallel design consistently outperforms the cascaded design in both clean and adversarial settings. By maintaining the independence of standard and equivariant properties, the parallel design achieves diverse and complementary feature representations that enhance adversarial resilience. Within the parallel design, simple concatenation of outputs consistently delivers higher robustness against FGSM and PGD attacks. This approach retains richer and more balanced feature representations, which are crucial for adversarial defense. Weighted sum fusion, while effective on clean data, is more vulnerable to adversarial attacks due to over-reliance on learned weights, which may fail to generalize under adversarial perturbations. The cascaded design exhibits diminished performance and robustness due to feature redundancy. The sequential dependence between standard and equivariant layers restricts the network's ability to capture diverse patterns, leading to reduced effectiveness in adversarial scenarios.

G Experiments Computing Resources

All experiments were performed on the Lawrence Supercomputer and NRP Nautilus HPC systems. Training and evaluation used a single GPU per experiment. Node specifications are provided in Table 5.

Table 5: Hardware configuration for experiments

Component	Configuration
CPUs	Dual 12-core SkyLake 5000 series
GPUs	Nvidia Tesla P100 16GB or Nvidia RTXA6000 64GB
RAM	64GB
SSD	240GB

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 claims made in the abstract and introduction are clearly aligned with the contributions and scope of the paper. The paper emphasizes enhancing adversarial robustness through equivariant architectures and demonstrates improved results on CIFAR-10, CIFAR-100, and CIFAR-10C datasets using group-equivariant convolutions. The claims are supported by both theoretical analysis and experimental results.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We have discussed the limitations of our methodology in Appendix B, where we reflect on potential areas for improvement and further exploration

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 theoretical analysis is clearly presented, with detailed assumptions and proofs, particularly regarding the relationship between equivariance and adversarial robustness. The paper provides a formal framework showing how equivariant architectures reduce hypothesis space complexity, regularize gradients, and tighten certified robustness bounds using the CLEVER framework.

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: The paper fully discloses the experimental setup, including the models, datasets, and evaluation strategies. It details the comparison of various models, the attack types used (FGSM and PGD), and experimental configurations, making it possible for others to replicate 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 will provide open access to the code. The datasets used (CIFAR-10, CIFAR-100, and CIFAR-10C) are publicly available. Once the code is shared, the experiments can be fully reproduced using these datasets. For CIFAR-10C, users will need to download it from an external source and update the data path accordingly. However, for CIFAR-10 and CIFAR-100, no additional setup is required—simply run the provided code.

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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are well described, including the split of the data set (CIFAR-10, CIFAR-100, CIFAR10C), the attack methods (FGSM and PGD), and the architecture details.

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: Error bars representing ±1 standard deviation over 5 random seeds are shown as shaded regions in Figures 1 and 2. Standard deviations are consistently below 1% across all models and experimental conditions (average std <0.8%).

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: In the appendix section of this paper, we have clearly outlined the hardware configuration used to run the experiments for the proposed methods.

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: The research adheres to the NeurIPS Code of Ethics, focusing on improving adversarial robustness in machine learning models without any known ethical issues related to data privacy, fairness, or security. There is no mention of potential ethical concerns or violations.

10. **Broader impacts**

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

    Answer: [NA]

    Justification: The paper focuses primarily on improving the adversarial robustness of AI models, with a specific emphasis on enhancing model resilience in challenging conditions. While the robustness of these models may have potential real-world applications, such as in medical science and autonomous vehicles, our work does not directly address the societal impacts of these applications. Therefore, the paper does not explore societal implications, and this question is marked as NA.

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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper focuses solely on enhancing the adversarial robustness of AI models to reduce mistakes in challenging conditions. As such, it does not involve the release of data or models that carry a high risk of misuse, and therefore, no specific safeguards are required

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: The datasets used in this paper (CIFAR-10, CIFAR-100, and CIFAR-10C) are publicly available. We have ensured that the terms of use for these datasets are respected, and the corresponding citations are included in the paper.

13. **New assets**

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

Answer: [No]

Justification: Our work does not create new assets such as datasets, however, we have shared our code.

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: [NA]

Justification: Our work does not involve crowdsourcing or research with human subjects. It focuses on improving adversarial robustness in AI models, and no human participants or crowdsourcing methods were used in the experiments.

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: [NA]

Justification: Our Work does not involve research with human subjects, and therefore, IRB approval or equivalent was not required. The focus of the paper is on improving adversarial robustness in AI models, which does not involve human participants.

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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not use large language models (LLMs) as part of the core methodology or research. LLMs were used only for editing and rephrasing text, such as improving grammar. This usage did not impact the scientific rigor or originality of the research.