Diverge to Converge: Mutual Heterogeneous Learning for Robust Pruning

Jinhui Yu[1,2], **Zikai Zhang**[3*], **Khaled A. Harras**[3], **Yidong Li**[1,2*]
[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation, Ministry of Education
[2]Beijing Jiaotong University
[3]Carnegie Mellon University

Abstract

Neural network pruning is crucial for efficient deployment on resource-constrained devices, yet achieving high sparsity often leads to significant robustness degradation against adversarial perturbations and corruptions. Recent works typically rely on single-model fine-tuning along a fixed optimization trajectory, which renders the network susceptible to local optima and noise while failing to restore the multiple robustness properties compromised during compression. In this paper, we propose **Mutual Heterogeneous Learning (MHL)**, a framework enabling robust pruning via **single-model inference**. **MHL** instantiates heterogeneity through two complementary mechanisms: **layer-wise Lipschitz regularization** for intermediate feature smoothness, and **adaptive margin objective** for difficulty-aware boundary separation. To guide these diverse experts to converge, we employ **entropy-based mutual distillation with a strategic schedule** that shifts the optimization trajectory from exploring diverse feature subspaces to consolidating a unified robust model. Extensive experiments on four clean and corruption benchmarks and adversarial attacks demonstrate that MHL significantly outperforms single-model baselines in both adversarial robustness (+5%) and corruption robustness (+2.6%), while maintaining competitive clean accuracy.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success across many vision tasks [Saeed *et al.*, 2017; Saeed *et al.*, 2019; Gao *et al.*, 2025], yet their deployment for various edge-based applications [Mostafa *et al.*, 2026; Mostafa *et al.*, 2024; Essameldin *et al.*, 2020; Abdelnasser *et al.*, 2020], on resource-constrained devices remains challenging due to substantial compute and memory demands [Fontanesi *et al.*, 2025; Zha *et al.*, 2025; Bartl *et al.*, 2025]. Structured pruning is a practical compression technique that removes redundant channels/filters to reduce latency and model size [Cheng *et al.*, 2024; Zhong *et al.*, 2025]. However, this efficiency gain is often achieved by sacrificing the network's resilience. [Piras *et al.*, 2025] As the model becomes leaner, it often suffers significant **robustness degradation**: they become substantially more vulnerable to both **adversarial perturbations** and **natural corruptions** than their unpruned counterparts. This trade-off poses a severe challenge for safety-critical applications, where efficiency cannot come at the expense of trustworthiness.

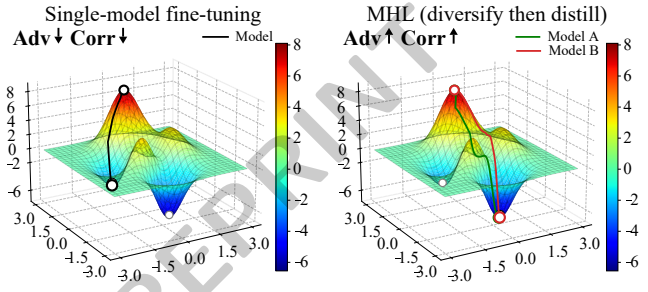

Figure 1: Convergence trajectory of fine-tuning on a single model and mutual heterogeneous learning on two models.

To mitigate this degradation, current approaches predominantly rely on adversarial training on pruned models, robustness-aware pruning, or specialized fine-tuning [Zhou *et al.*, 2025; Saxena *et al.*, 2024; Zhang *et al.*, 2023]. However, we argue that these methods fundamentally follow a **single-model paradigm**, which is limited by their dependence on a single optimization trajectory and struggles to simultaneously recover multiple robustness properties after compression. As illustrated in Figure 1(a), fine-tuning on non-convex landscapes restricts a lone network to a single trajectory, often resulting in sharp local minima. Crucially, such constrained optimization on one pruned model fails to simultaneously maximize both internal feature smoothness and external boundary separation, inevitably yielding a suboptimal trade-off and recovery in multiple robustness properties.

To bridge the gap between efficiency and robustness, we propose a paradigm shift from single-trajectory optimization to a collaborative mutual learning framework. Our key insight is *diverge to converge*, instead of forcing one pruned model to be universally robust, we recover robustness through **designed complementarity**. As visualized in Figure 1(b), we instantiate two heterogeneous models driven by comple-

mentary objectives. For "diverge", two pruned models are intentionally driven toward different robustness behaviors, their failure regions can become less correlated. For "converge", with proper mutual supervision, they can collaboratively cover each other's vulnerabilities and yield stronger robustness. This strategy effectively expands the search space and synthesizes multifaceted robustness into a single model without incurring extra inference costs.

Concretely, we propose **Mutual Heterogeneous Learning (MHL)**, a dual-model co-training framework for robust pruning with **single-model inference**, strictly adhering to edge application constraints. For "diverge" in MHL, heterogeneous objectives have two complementary mechanisms: **layer-wise Lipschitz regularization** (feature smoothness under local perturbations) to enforce intermediate feature smoothness, and an **adaptive margin objective** (difficulty-aware logit margin) for difficulty-aware boundary separation. Crucially, this design bridges the gap between *internal feature stability* and *external decision separability*, ensuring that the pruned model is not only insensitive to input noise but also decisive in its classifications. For "converge" in MHL, we employ **entropy-based mutual distillation under a dynamic schedule**. By adaptively regulating the optimization focus across training stages, this strategy actively shifts the trajectory from initially exploring diverse feature subspaces to consolidating a unified consensus, thereby restoring multiple robustness properties while strictly maintaining inference efficiency by **deploying only a single pruned model** at test time.

We conduct extensive experiments under structured pruning on four benchmarks, evaluating clean accuracy (CIFAR-10/100), adversarial robustness (FGSM/PGD/PGD-20), and corruption robustness (CIFAR-10/100-C). Our results show MHL outperforming post-pruning baselines across robustness metrics while maintaining competitive clean accuracy. Ablation studies further validate the necessity of heterogeneity, mutual distillation, and feature diversity regularization.

Our main contributions are summarized as follows:

- We propose **MHL**, a novel structured pruning framework that shifts the robust pruning paradigm from single-trajectory fine-tuning to a "diverge-to-converge" strategy. This approach avoids the dependency on a single optimization trajectory, mitigates local optima, and synthesizes a robust consensus from diverse optimization directions.
- We instantiate heterogeneity via two mechanisms: layer-wise Lipschitz regularization to enforce internal feature stability, and an adaptive margin objective to ensure external decision separability. We synthesized these via entropy-based mutual distillation and feature diversity regularization that dynamically transitions from subspace exploration to robust consensus.
- We demonstrate consistent improvements on adversarial robustness and corruption robustness under structured pruning, with ablations confirming the role of each component. MHL improves adversarial robustness by an average of 5% and corruption robustness by 2.6% over prior methods.

2 Related Works

2.1 Adversarial Training Methods

Adversarial Training (AT) [Madry *et al.*, 2018] is the most effective defense against adversarial examples [Goodfellow *et al.*, 2014]. To improve the robustness-accuracy trade-off, early variants like TRADES [Zhang *et al.*, 2019] and MART [Wang *et al.*, 2019] introduced regularization-based objectives. More recently, the focus has shifted to feature-space manipulations, such as explicit feature disentanglement in AFD [Zhou *et al.*, 2025] and asymmetric representation regularization in AR-AT [Waseda *et al.*, 2025]. Despite their effectiveness, these methods train dense models, which limits their deployment efficiency. Our approach extends adversarial robustness to the pruned model regime through mutual heterogeneous learning.

2.2 Robust Neural Network Pruning

Standard pruning strategies [Han *et al.*, 2015; Filters'Importance, 2016; Liu *et al.*, 2017; Frankle and Carlin, 2019], though effective for clean accuracy, suffer from severe robustness degradation [Ye *et al.*, 2019; Gui *et al.*, 2019]. To address this, robustness-aware pruning integrates adversarial objectives into the compression process. Prominent approaches focus on learning robust importance scores or saliency rankings [Sehwag *et al.*, 2020; Zhao and Wressnegger, 2024; Lee *et al.*, 2022]. Alternatively, some works explore dynamic sparse training [Chen *et al.*, 2022] or weight reparameterization [Li *et al.*, 2023] to mitigate robust overfitting and maintain representational capacity. However, these methods strictly adhere to a single-model paradigm, limiting their ability to escape sharp local optima or recover multi-faceted robustness. Our MHL overcomes this via heterogeneous mutual learning, effectively synthesizing complementary robustness properties without incurring additional inference costs.

2.3 Ensemble Methods and Architectural Defenses

Ensemble methods leverage diversity to mitigate attack transferability [Kariyappa and Qureshi, 2019; Pang *et al.*, 2019]. Recent approaches maximize this via explicit diversity regularization [Yang *et al.*, 2020; Jung and Song, 2025] or layer-wise orthogonalization [Ebrahimpour Boroojeny *et al.*, 2025]. However, these strategies typically incur prohibitive inference costs. Parallel efforts explore architectural modifications, such as input filtering layers [Suresh *et al.*, 2025] or neuron-level monosemanticity constraints [Zhang *et al.*, 2024]. While promising, these architectural defenses are often specific to certain backbones and lack generalizability.

Unlike ensembles that multiply inference costs, MHL adopts a *Two-model training, single-model deployment* strategy. By distilling heterogeneous experts via a **diverge-to-converge** trajectory, we condense multifaceted robustness into a single pruned model, achieving ensemble-level resilience with zero additional overhead.

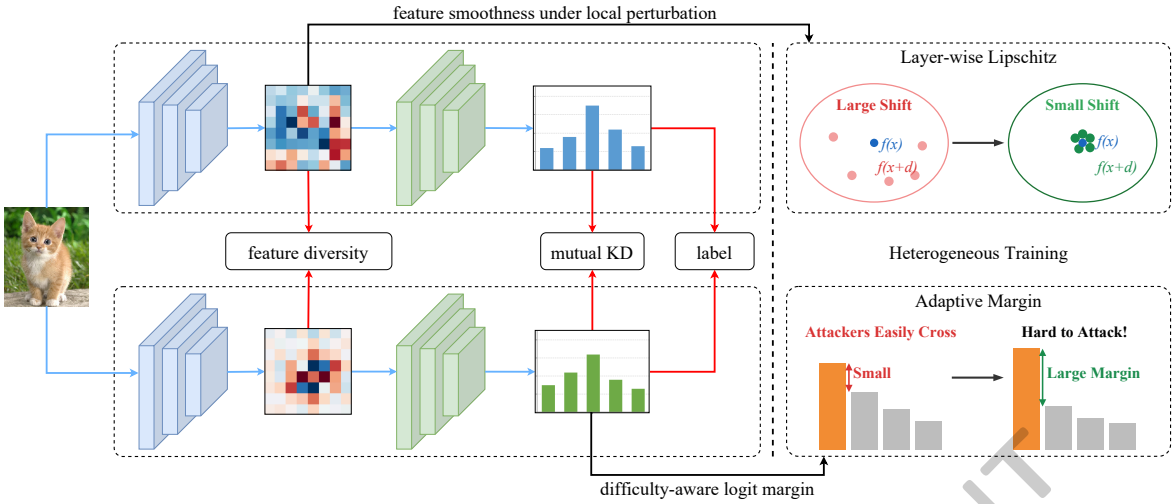

Figure 2: Overview of MHL. Two branches are pruned and trained jointly with heterogeneous objectives (*Lipschitz vs. Margin*) and synchronized via *entropy-gated mutual distillation*. A *feature diversity* loss prevents collapse. **Only one model is retained for inference**.

3 Method

3.1 Preliminaries and Problem Setting

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a labeled dataset with inputs $x \in \mathbb{R}^d$ and labels $y \in \{1, \dots, C\}$. A classifier $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$ parameterized by θ produces logits $z = f_\theta(x)$, and the predictive distribution is $p = \mathrm{softmax}(z)$. We consider **structured pruning**—removing entire filters or channels—which yields models directly accelerable on commodity hardware without specialized sparse libraries.

Three-stage pipeline. Following the standard pipeline (Pre-train $\to$ Prune $\to$ Fine-tune), our work focuses exclusively on the final stage: given a pruned checkpoint $\tilde{f}_\theta = \mathcal{P}(f_\theta)$, we aim to maximize its adversarial and corruption robustness without increasing inference cost.

3.2 Mutual Heterogeneous Learning (MHL)

We propose **Mutual Heterogeneous Learning (MHL)**, a co-training framework that recovers robustness of pruned models through *designed complementarity*. The core insight is *diverge to converge*: rather than forcing a single compact model to satisfy all robustness requirements simultaneously, we train two pruned models with deliberately different inductive biases, then leverage their complementary strengths through mutual supervision.

Two-model mutual training, single-model deployment. Starting from the same pruned-and-finetuned checkpoint $\tilde{f}_\theta$, we create two branches with identical architecture: $\tilde{f}_{\theta_1}$ and $\tilde{f}_{\theta_2}$. Both models share the same pruned structure (e.g., 50% filters removed) but diverge during training due to heterogeneous regularization. Crucially, *only one model is retained at deployment*—the better-performing branch on a held-out validation set—thus preserving the efficiency benefits of pruning with **zero additional inference cost**.

Base classification objective. Both models are supervised with ground-truth labels via cross-entropy:

$$\mathcal{L}_{\mathrm{CE}}^{(i)} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\log p_y^{(i)}\right], \quad p^{(i)} = \mathrm{softmax}(\tilde{f}_{\theta_i}(x)). \tag{1}$$

MHL augments $\mathcal{L}_{\mathrm{CE}}^{(1)} + \mathcal{L}_{\mathrm{CE}}^{(2)}$ with three synergistic components: **(i)** heterogeneous constraints to induce complementary robustness behaviors, **(ii)** entropy-gated mutual distillation to transfer reliable knowledge, and **(iii)** feature diversity regularization to prevent representation collapse.

3.3 Heterogeneous Constraints

The key to MHL is imposing *different* regularizers on the two branches: **layer-wise Lipschitz** on $\tilde{f}_{\theta_1}$ (encouraging smooth internal representations) and **adaptive margin** on $\tilde{f}_{\theta_2}$ (enforcing confident logit separation). This asymmetric design is intentional—the two constraints target orthogonal aspects of robustness, producing complementary failure modes that can be exploited via mutual learning.

Layer-wise Lipschitz Regularization

Adversarial perturbations exploit the sensitivity of neural network activations. We encourage $\tilde{f}_{\theta_1}$ to have stable intermediate features under small input perturbations. Let $\phi_\ell^{(1)}(x)$ denote the feature map at stage ℓ of a four-stage ResNet ($\ell \in \{1, 2, 3, 4\}$). For each input x, we sample K isotropic Gaussian perturbations $\delta_k \sim \mathcal{N}(0, \rho^2 I)$ and penalize feature displacement across all layers:

$$\mathcal{L}_{\mathrm{LL}} = \mathbb{E}_x\left[\frac{1}{K}\sum_{k=1}^{K}\sum_{\ell=1}^{4} w_\ell \left\|\phi_\ell^{(1)}(x+\delta_k) - \phi_\ell^{(1)}(x)\right\|_2\right], \tag{2}$$

where $w_\ell = (0.25, 0.25, 0.25, 0.25)$ are equal layer weights, $K = 3$ samples, and $\rho = 0.1$ is the perturbation radius. Importantly, the center features $\phi_\ell^{(1)}(x)$ are computed *without gradients* (stop-gradient), so only the perturbed path contributes to the regularization. Compared to end-to-end Lipschitz bounds [Cissé *et al.*, 2017], layer-wise control offers

finer granularity, particularly valuable for pruned networks where reduced redundancy amplifies sensitivity.

Adaptive Margin Regularization

To complement the representation-level smoothness of $\tilde{f}_{\theta_1}$, we enforce *decision-level* robustness on $\tilde{f}_{\theta_2}$ via logit margin. Given logits $z^{(2)} = \tilde{f}_{\theta_2}(x)$, let z_y be the correct-class logit and $z_{\max} = \max_{c \neq y} z_c^{(2)}$ the largest incorrect logit. We define sample difficulty as

$$d(x) = 1 - p_y^{(2)}, \quad p^{(2)} = \mathrm{softmax}(z^{(2)}), \tag{3}$$

and set a sample-adaptive target margin:

$$m(x) = m_0 \left(1 - \tfrac{1}{2} d(x)\right), \tag{4}$$

where m_0 is a base margin hyperparameter (default $m_0 = 10$). The loss penalizes insufficient margins:

$$\mathcal{L}_{\mathrm{AM}} = \mathbb{E}_{(x,y)} \left[\max\big(0,\ m(x) - (z_y - z_{\max})\big)\right]. \tag{5}$$

This adaptive design imposes larger margins on easy (high-confidence) samples while relaxing constraints on hard ones, avoiding gradient explosion on inherently ambiguous inputs and stabilizing optimization.

3.4 Entropy-Gated Mutual Distillation

Heterogeneous constraints alone may lead to excessive divergence without transferring complementary knowledge. We bridge the two branches via *entropy-gated mutual distillation*, where the more confident model on each sample acts as the teacher.

For model i, we measure prediction uncertainty via entropy:

$$H(p^{(i)}) = -\sum_{c=1}^{C} p_c^{(i)} \log p_c^{(i)}. \tag{6}$$

Lower entropy indicates higher confidence. We define binary teaching indicators: $\mathbb{I}_{1\to 2} = \mathbb{I}[H(p^{(1)}) < H(p^{(2)})]$ and $\mathbb{I}_{2\to 1} = 1 - \mathbb{I}_{1\to 2}$. Using temperature τ (default $\tau = 3$) and softened probabilities $p_\tau^{(i)} = \mathrm{softmax}(z^{(i)}/\tau)$, the distillation losses are:

$$\mathcal{L}_{\mathrm{KD}}^{2\leftarrow 1} = \mathbb{E}_x \left[\mathbb{I}_{1\to 2} \cdot \tau^2 \,\mathrm{KL}\Big(p_\tau^{(1)} \,\|\, p_\tau^{(2)}\Big)\right], \tag{7}$$

$$\mathcal{L}_{\mathrm{KD}}^{1\leftarrow 2} = \mathbb{E}_x \left[\mathbb{I}_{2\to 1} \cdot \tau^2 \,\mathrm{KL}\Big(p_\tau^{(2)} \,\|\, p_\tau^{(1)}\Big)\right]. \tag{8}$$

The teacher's logits are *detached* (stop-gradient) to prevent symmetric updates. The total distillation loss is $\mathcal{L}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{KD}}^{1\leftarrow 2} + \mathcal{L}_{\mathrm{KD}}^{2\leftarrow 1}$.

This sample-wise gating mechanism differs from standard mutual learning [Zhang *et al.*, 2018], which transfers knowledge bidirectionally on all samples. Our entropy gating ensures that only *reliable* soft targets are transferred, reducing error propagation and stabilizing co-training dynamics.

3.5 Feature Diversity Regularization

Mutual distillation encourages output agreement, but may inadvertently collapse the two branches to identical internal representations, defeating the purpose of heterogeneous training. We introduce a *feature diversity loss* to explicitly maintain representational differences.

For each stage ℓ, we extract global-average-pooled features $\bar{\phi}_\ell^{(i)}(x) \in \mathbb{R}^{d_\ell}$ and ℓ_2-normalize them to unit vectors $\hat{\phi}_\ell^{(i)}(x)$. The diversity loss penalizes high cosine similarity:

$$\mathcal{L}_{\mathrm{div}} = \mathbb{E}_x \left[\sum_{\ell=1}^{4} \alpha_\ell \left\langle \hat{\phi}_\ell^{(1)}(x),\, \hat{\phi}_\ell^{(2)}(x) \right\rangle\right], \tag{9}$$

where $\alpha_\ell = (0.5, 1.0, 1.0, 0.5)$ are layer weights that emphasize middle layers (which capture most semantic content). Minimizing $\mathcal{L}_{\mathrm{div}}$ encourages *orthogonal* feature directions between the two branches, preserving complementarity throughout training.

3.6 Overall Objective and Optimization

Combining all components, the MHL training objective is:

$$\mathcal{L}_{\mathrm{MHL}} = \underbrace{\mathcal{L}_{\mathrm{CE}}^{(1)} + \mathcal{L}_{\mathrm{CE}}^{(2)}}_{\text{supervised}} + \underbrace{\lambda_{\mathrm{LL}}\mathcal{L}_{\mathrm{LL}} + \lambda_{\mathrm{AM}}\mathcal{L}_{\mathrm{AM}}}_{\text{heterogeneous}} + \underbrace{\lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}}_{\text{mutual}} + \underbrace{\lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}}_{\text{diversity}}. \tag{10}$$

We optimize Eq. (10) end-to-end using SGD with momentum (0.9) and cosine annealing learning rate schedule. Gradient clipping (max norm 1.0) is applied to both branches for training stability. Default hyperparameters are $\lambda_{\mathrm{LL}} = \lambda_{\mathrm{AM}} = 0.1$, $\lambda_{\mathrm{KD}} = 0.5$, $\lambda_{\mathrm{div}} = 0.1$.

The optimization concludes by discarding the auxiliary branch and retaining only the primary pruned model for deployment, ensuring zero additional inference cost compared to standard pruning.

Design rationale. The four loss terms form a balanced system: **(i)** $\mathcal{L}_{\mathrm{LL}}$ and $\mathcal{L}_{\mathrm{AM}}$ induce complementary inductive biases—one enforces *representation smoothness*, the other enforces *decision confidence*; **(ii)** $\mathcal{L}_{\mathrm{KD}}$ transfers reliable knowledge from the stronger model to the weaker one on each sample; **(iii)** $\mathcal{L}_{\mathrm{div}}$ prevents the distillation from collapsing both models to identical solutions.

3.7 Analysis of Robustness Components

We analyze why the two heterogeneous training components in our implementation (*layer-wise Lipschitz constraints* and *adaptive margin*) are complementary, and how mutual supervision further enlarges the robust radius of each pruned model. **Crucially, while this analysis establishes the static theoretical lower bound, our dynamic scheduling which is driven by the feature diversity loss, serves as the practical mechanism to prevent representational collapse, ensuring that the final model effectively approximates this optimal robustness limit**.

Theoretical Motivation and Instantiation. The certifiable robust radius of a classifier is lower-bounded by the ratio of its decision margin to its Lipschitz constant: $r_f(x) \geq \frac{m_f(x)}{L}$ [Tsuzuku *et al.*, 2018]. This inequality reveals two complementary levers for robustness enhancement: increasing the numerator $m_f(x)$ (via boundary separation) and decreasing the denominator L (via feature smoothness). To explicitly target these complementary levers, we instantiate MHL by training two pruned models $\{f^{(1)}, f^{(2)}\}$ jointly with

Method	25% pruning				50% pruning				75% pruning			
	Clean	FGSM	PGD-10	PGD-20	Clean	FGSM	PGD-10	PGD-20	Clean	FGSM	PGD-10	PGD-20
FT	94.22	65.79	55.82	56.20	93.61	61.10	50.76	51.29	88.96	38.5	36.74	36.77
AT	86.21	82.58	82.44	82.45	84.32	80.19	80.11	80.08	79.81	75.39	75.34	75.35
SWA	95.14	66.22	60.31	60.82	94.54	62.29	57.64	58.29	91.47	48.21	46.15	46.47
ASAM	95.08	65.46	61.74	62.17	94.48	62.12	58.92	59.16	91.47	46.42	45.28	45.73
AdaSAP	94.28	67.08	55.66	56.26	93.70	61.24	51.70	52.11	89.16	39.99	36.99	37.24
AFD	80.45	77.57	77.54	77.54	76.63	73.55	73.51	75.50	70.11	67.19	67.15	67.15
EED	83.28	77.39	77.24	77.25	74.68	69.67	69.57	69.58	64.63	60.18	60.17	60.18
ANF	83.45	63.28	61.21	61.24	83.70	62.26	60.10	60.11	82.15	60.03	58.04	58.00
LOTOS	95.00	68.40	64.56	64.84	94.43	62.22	58.66	58.87	91.36	54.76	51.97	52.55
MONO	95.14	73.07	63.05	62.90	94.59	70.51	60.48	60.28	91.86	56.06	45.83	46.35
MHL (ours)	92.99	77.38	71.59	71.54	92.81	76.88	71.08	71.13	89.57	61.70	55.29	55.52

Table 1: Main results on CIFAR-10 under structured pruning with single-model inference.

heterogeneous objectives. Specifically, we assign distinct regularizers to each model, one focusing on **Adaptive Margin** to maximize the numerator, and the other on **Layer-wise Lipschitz Regularization** to minimize the denominator. So, we can decouple the complex robustness objective into two manageable sub-problems.

Layer-wise Lipschitz constraints as local smoothing (denominator control) Formally, for a network $f(x) = W_l\phi(...W_1x)$, the global Lipschitz constant L is upper-bounded by the product of layer-wise spectral norms: $L \leq \prod_{i=1}^{l} \|W_i\|_2$.

Proposition 1: *Minimizing the layer-wise spectral norms explicitly decreases the global Lipschitz constant L. According to the robust radius bound $r_f(x) \geq m_f(x)/L$, this operation effectively performs "denominator control," enlarging the certified robust radius by suppressing feature amplification.*

Mutual supervision transfers the two gains across models. While Layer-wise Lipschitz constraints create diverse models (one with high margin m_f, one with low Lipschitz L), we need to unify these gains. We employ an entropy-based mutual distillation loss $\mathcal{L}_{KD}$.

Proposition 2 (Robustness Transfer). *We employ an entropy-gated mutual distillation loss $\mathcal{L}_{KD}$ that dynamically designates the lower-entropy agent as the teacher. By minimizing the KL divergence $KL(p^{(student)}||p^{(teacher)})$ directionally, MHL explicitly transfers the superior geometric property (e.g., larger margin or smoother features) from the stronger agent to the weaker one, thereby enlarging the robust radius of the ensemble consensus.*

Synergy: robust radius of the smooth branch increases beyond using a single component. Finally, we analyze how the interaction between the two agents surpasses the limits of independent training. While the smooth branch minimizes the denominator L, it risks collapsing the numerator m_f (decision margin) due to over-regularization.

Proposition 3 (Synergy of Heterogeneity). *Let $r_{base}^{(1)}$ be the certified robust radius of $f^{(1)}$ when trained in isolation. Under MHL, the effective radius of the smooth agent $r_{MHL}^{(1)}$ satisfies:*

$$r_{MHL}^{(1)}(x) \geq r_{base}^{(1)}(x) + \underbrace{\frac{\Delta m_{2\to 1}}{L^{(1)}}}_{\text{Gain from } f^{(2)}} \tag{11}$$

where $\Delta m_{2\to 1} > 0$ represents the margin gain transferred from the margin-aware teacher $f^{(2)}$ via entropy-gated mutual distillation, and $L^{(1)}$ is the low Lipschitz constant explicitly maintained by $f^{(1)}$.

Remark 1: This inequality mathematically formalizes the "Diverge-to-Converge" benefit. Crucially, the **explicit Lipschitz constraint** on $f^{(1)}$ acts as a geometric anchor, ensuring the denominator $L^{(1)}$ remains suppressed even as the numerator increases via distillation. This allows the smooth agent to "borrow" the margin advantage (Δm) from $f^{(2)}$ without compromising its own stability, achieving a robust radius impossible for either model in isolation.

Remark 2 (Diverge-Converge Principle). Simultaneously maximizing the margin m (numerator) and minimizing the Lipschitz constant L (denominator) is often an ill-posed optimization problem for a single sparse network with limited capacity. MHL circumvents this by *decoupling* these conflicting objectives across heterogeneous agents and subsequently *aggregating* their specific advantages (high separability from $f^{(2)}$ and high smoothness from $f^{(1)}$) via mutual supervision, effectively realizing the theoretical synergy described in Proposition 2.

4 Experiments

Datasets. Following the edge application constraint in [Ebrahimpour Boroojeny *et al.*, 2025], we evaluate the performance of MHL on CIFAR-10 and CIFAR-100, including clean and adversarial samples. For corruption robustness, we also evaluate CIFAR-10-C and CIFAR-100-C, and illustrate category-wise and average accuracy.

Implementation Details. All methods follow the same pipeline: dense models are pretrained first, followed by structured pruning and fine-tuning for 20 epochs to obtain a common initialization, and finally trained for 50 epochs using MHL or a baseline training strategy. We consider three structured pruning rates: 25%, 50%, and 75%. All methods at

Method	25% pruning				50% pruning				75% pruning			
	Clean	FGSM	PGD-10	PGD-20	Clean	FGSM	PGD-10	PGD-20	Clean	FGSM	PGD-10	PGD-20
FT	74.71	27.91	25.1	25.32	72.38	21.81	19.46	19.43	63.55	16.05	14.06	13.88
AT	61.26	54.31	54.10	54.09	60.49	53.99	53.89	53.90	53.71	48.72	48.64	48.67
SWA	75.81	33.62	30.80	31.00	74.25	28.63	26.75	26.63	68.73	22.51	20.80	20.93
ASAM	75.94	33.69	30.88	31.00	73.68	27.07	25.35	25.35	68.01	20.28	18.13	18.35
AdaSAP	74.67	31.14	28.08	27.91	72.48	25.87	23.41	23.26	63.67	18.65	16.87	16.73
AFD	55.71	51.81	51.78	51.78	52.41	48.61	48.55	48.56	44.23	41.38	41.34	41.35
EED	55.97	49.85	49.69	49.74	50.13	42.89	42.68	42.71	36.19	32.02	31.99	31.98
ANF	85.11	64.44	62.51	62.49	83.34	61.96	60.04	60.09	81.85	57.33	55.94	56.11
LOTOS	78.17	37.18	34.04	34.44	76.57	33.82	31.55	31.37	69.77	28.57	26.70	26.65
MONO	75.60	35.36	35.60	35.42	73.91	29.23	27.50	27.37	67.54	19.94	17.44	17.31
MHL (ours)	73.66	43.13	38.75	38.74	71.36	36.55	32.70	32.63	63.41	30.99	28.19	28.39

Table 2: Main results on CIFAR-100 under structured pruning with single-model inference.

Method	25% pruning					50% pruning					75% pruning				
	Avg	Noise	Blur	Weather	Digital	Avg	Noise	Blur	Weather	Digital	Avg	Noise	Blur	Weather	Digital
FT	73.75	54.60	71.57	84.50	81.15	71.44	51.22	69.36	82.53	79.27	64.35	40.91	63.08	75.97	73.28
AT	77.94	80.08	79.90	75.78	76.31	76.47	79.11	78.00	74.77	74.57	72.24	75.96	73.81	69.20	70.47
SWA	75.93	56.14	74.88	86.22	83.14	73.77	51.58	73.01	85.04	81.66	68.01	44.49	66.50	79.91	77.03
ASAM	75.32	56.95	72.78	85.66	82.79	73.77	53.17	72.41	84.81	81.28	67.93	46.37	65.48	79.61	76.57
AdaSAP	74.09	55.38	72.26	84.60	81.07	69.63	54.09	59.04	83.89	79.31	64.49	42.09	63.06	75.94	72.97
AFD	72.86	76.15	74.54	69.97	71.20	69.44	73.15	71.17	65.80	67.95	63.76	67.75	65.74	59.14	62.54
EED	73.29	80.85	71.26	70.67	71.69	66.85	72.66	68.00	61.42	65.63	56.41	62.60	56.40	51.08	55.83
ANF	72.99	74.17	70.29	74.36	73.55	72.08	72.82	69.69	73.41	72.70	69.70	68.87	66.10	72.80	71.20
LOTOS	74.57	55.65	72.05	85.11	82.25	72.35	53.79	69.39	83.04	80.06	68.16	49.43	65.06	77.78	76.82
MONO	73.70	53.62	70.20	85.50	82.13	69.22	51.53	57.67	83.96	80.82	66.22	42.27	63.62	78.88	75.90
MHL (ours)	77.27	71.31	71.73	84.97	80.71	76.84	68.87	73.60	83.85	80.19	69.71	58.33	65.44	78.83	74.79

Table 3: Corruption robustness on CIFAR-10-C. We report accuracy (%) for Avg and four corruption categories.

each pruning rate share the same initialization. We use the ℓ_∞ threat model with perturbation budget $\epsilon = 1/255$ and report clean, FGSM, PGD-10/20 accuracy.

4.1 Baselines

We compare against robust training and pruning-related baselines, all the methods are implemented in the final training phase, same as our MHL. The baselines include vanilla fine-tuning (FT), AT [Madry *et al.*, 2018], SWA, ASAM [Kwon *et al.*, 2021], AdaSAP [Bair *et al.*, 2023], AFD [Zhou *et al.*, 2025], EED [Jung and Song, 2025], ANF [Suresh *et al.*, 2025], LOTOS [Ebrahimpour Boroojeny *et al.*, 2025], and MONO [Zhang *et al.*, 2024]. For methods originally designed for ensembles or different training regimes, we use their adapted variants under the same pipeline. Most baseline use single-model inference, for ensemble methods (EED, LOTOS), we set 3 sub-models for their evaluation.

4.2 Main Results on CIFAR-10/100

Results on CIFAR-10. Table 1 shows that MHL consistently improves adversarial robustness of pruned models while keeping clean accuracy competitive. At 50% pruning, MHL boosts PGD-20 accuracy from 51.29% (FT) and 60.28% (MONO) to 71.13%, indicating that complementary constraints plus mutual learning recover robustness properties that single-model fine-tuning fails to restore. Compared to optimization-centric baselines (SWA/ASAM), MHL yields markedly higher PGD robustness. While adversarial training (AT) achieves stronger attack accuracy, it incurs a substantial clean-accuracy drop, whereas MHL retains high clean performance (92.81%) and delivers a stronger robustness–accuracy trade-off under structured pruning. Under more aggressive compression, MHL still improves PGD-20 over FT by a large margin, demonstrating robustness recovery in the most challenging regime.

Results on CIFAR-100. Table 2 further confirms the benefit of MHL on a harder dataset. At 50% pruning, MHL improves PGD-20 from 19.43% (FT) and 27.37% (MONO) to 32.63%, and at 75% pruning it more than doubles FT's PGD-20 accuracy. These gains are achieved with substantially higher clean accuracy than AT, suggesting that MHL is particularly effective for robustness recovery when pruning amplifies capacity constraints.

4.3 Corruption Robustness on CIFAR

Results on CIFAR-10-C. Table 3 shows that MHL achieves strong average corruption robustness across pruning rates and exhibits clear category-wise complementarity. In particular, MHL substantially improves robustness on Weather and Digital corruptions, which aligns with our design that combines feature-space smoothing and decision-boundary shaping. Although some baselines (e.g., AT) can be stronger on Noise and Blur, MHL maintains competitive averages and better overall balance when considering both adversarial and cor-

Method	25% pruning					50% pruning					75% pruning				
	Avg	Noise	Blur	Weather	Digital	Avg	Noise	Blur	Weather	Digital	Avg	Noise	Blur	Weather	Digital
FT	45.93	22.79	47.49	56.89	52.77	42.98	21.47	44.11	53.70	49.23	37.22	16.77	39.40	46.00	43.20
AT	46.55	42.73	50.66	44.30	47.19	46.69	45.24	49.64	44.38	46.76	43.16	42.62	46.36	40.45	42.66
SWA	47.62	23.29	49.22	58.73	55.11	45.50	22.63	46.74	56.64	52.30	41.64	19.26	43.89	51.45	48.15
ASAM	47.21	24.01	48.29	58.25	54.43	44.25	21.36	45.63	54.97	51.21	40.29	18.68	41.99	49.61	47.07
AdaSAP	45.72	22.46	46.99	56.73	52.85	43.04	21.32	44.49	53.78	49.15	37.01	16.48	39.70	45.42	42.86
AFD	45.94	44.93	49.24	43.41	45.55	43.05	42.59	46.22	40.25	42.59	36.78	38.01	39.56	33.45	35.87
EED	45.18	47.46	46.98	42.99	43.63	39.54	40.85	41.66	37.25	38.42	28.69	32.08	30.53	24.58	27.65
LOTOS	48.34	25.64	48.49	59.69	55.79	46.72	23.63	47.73	57.25	54.23	42.86	22.95	44.31	51.41	49.23
MONO	44.69	20.76	45.45	56.14	52.39	42.27	19.25	42.78	54.16	49.27	38.73	16.41	40.72	48.21	45.65
MHL (ours)	48.74	33.00	48.27	57.33	53.90	45.64	30.02	44.67	55.13	50.52	41.00	26.42	40.93	49.08	45.40

Table 4: Corruption robustness on CIFAR-100-C. We report accuracy (%) for Avg and four corruption categories.

Pruning method	25%			50%			75%		
	Clean	Corrupt	PGD	Clean	Corrupt	PGD	Clean	Corrupt	PGD
Magnitude-L1	92.99	77.27	71.54	92.81	76.84	71.13	89.57	69.71	55.52
Magnitude-L2	93.03	76.99	67.96	92.71	76.73	71.37	90.13	70.68	55.02
Taylor	93.33	78.18	74.90	92.76	75.86	70.89	89.80	70.32	56.76

Table 5: Ablation on structured pruning methods. "Corrupt" denotes Avg accuracy on CIFAR-10-C; "PGD" denotes PGD-20 with $\epsilon = 1/255$.

Method	Clean	Corrupt	FGSM	PGD-10	PGD-20
entropy	92.81	76.84	76.88	71.08	71.13
confidence	92.79	76.24	76.14	70.20	70.36
consistency	90.99	75.29	67.73	63.59	63.77
alignment	91.89	75.01	75.84	69.05	70.71
disagreement	91.76	78.51	76.27	60.60	59.60

Table 6: Ablation on mutual learning scheme and feature diversity loss (CIFAR-10, 50% pruning).

ruption robustness under single-model deployment.

Results on CIFAR-100-C. As shown in Table 4, MHL yields consistent improvements over fine-tuning and robustness recovery baselines, especially under heavier pruning. For instance, at 75% pruning, MHL improves the average corruption accuracy by +3.78 points over FT (41.00 vs. 37.22), demonstrating that heterogeneous co-training remains beneficial when the pruned model's robustness is severely degraded.

4.4 Ablation Studies

Analysis of pruning methods. Table 5 shows that MHL is robust to the choice of structured pruning criterion; all three pruning rules yield similar clean and corruption accuracy, indicating that MHL is the dominant factor for robustness recovery. Meanwhile, the pruning criterion can affect adversarial robustness under heavy compression: at 25% pruning, Taylor pruning yields higher PGD-20 than Magnitude-L1 and Magnitude-L2; at 75% pruning, Taylor again achieves the best PGD-20 while maintaining strong corruption robustness. Given this mild sensitivity, we adopt Magnitude-L1 as the default for its simplicity and stability, and emphasize that MHL consistently improves robustness across pruning protocols.

Analysis of mutual learning schemes. Table 6 validates that how we couple the heterogeneous branches is critical. Our default entropy-gated distillation achieves the best overall balance (Clean 92.81%, Corrupt 76.84%, PGD-20 71.13%), outperforming alternative couplings such as confidence gating and alignment. Notably, disagreement-driven coupling can boost corruption robustness (78.51%) but severely hurts adversarial robustness, suggesting that encouraging output mismatch may increase robustness to distribution shift while weakening worst-case robustness. In contrast, consistency coupling degrades both clean and attack performance, indicating that overly strong agreement can suppress the benefits of heterogeneity. These results suggest that entropy gating transfers only reliable soft targets, enabling effective knowledge sharing while preserving complementary robustness behaviors.

5 Conclusion

In this work, we identified the reliance on a *single optimization trajectory* as the fundamental bottleneck in current robust pruning methods, which inevitably leads to local minima and suboptimal robustness trade-offs. To bridge this limitation, we proposed Mutual Heterogeneous Learning (MHL), a novel framework governed by the core philosophy of "diverge to converge". By orchestrating heterogeneous models to explicitly decouple robustness into *internal feature stability* and *external decision separability*, and subsequently synthesizing them via dynamic consensus, MHL successfully recovers multiple robustness properties. Extensive experiments on 4 benchmarks and under diverse adversarial attacks demonstrate that MHL consistently outperforms strong single-model baselines, yielding substantial gains in both adversarial robustness (+5%) and corruption robustness (+2.6%) while maintaining competitive clean accuracy. Crucially, our approach demonstrates that leveraging *mutual heterogeneity learning* is a potent strategy to enhance the trustworthiness of compact models while maintaining single-model inference efficiency.

Acknowledgments

Zikai Zhang (zikaiz2@cmu.edu) and Yidong Li (ydli@bjtu.edu.cn) are the corresponding authors. This work was supported in part by ARG02-0429-240389 from the Qatar National Research Fund/Qatar Research, Development and Innovation Council; the National Science Foundation of China under Grant U2268203, 62306027, 62402031 and 62372436; the Beijing Natural Science Foundation 4264109; the ZTE Corporation project Sim-to-Real Technology Exploration for Robotics (K25L01090); and the Beijing Nova Program (No. 20240484620).

References

[Abdelnasser *et al.*, 2020] Heba Abdelnasser, Khaled A Harras, and Moustafa Youssef. Magstroke: A magnetic based virtual keyboard for off-the-shelf smart devices. In *IEEE SECON*, pages 1–9, 2020.

[Bair *et al.*, 2023] Anna Bair, Hongxu Yin, Maying Shen, Pavlo Molchanov, and Jose Alvarez. Adaptive sharpness-aware pruning for robust sparse networks. *arXiv preprint arXiv:2306.14306*, 2023.

[Bartl *et al.*, 2025] Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6):1–36, 2025.

[Chen *et al.*, 2022] Tianlong Chen, Zhenyu Zhang, Pengjun Wang, Santosh Balachandra, Haoyu Ma, Zehao Wang, and Zhangyang Wang. Sparsity winning twice: Better robust generalization from more efficient training. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2202.09844.

[Cheng *et al.*, 2024] Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Cissé *et al.*, 2017] Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 854–863, 2017.

[Ebrahimpour Boroojeny *et al.*, 2025] Ali Ebrahimpour Boroojeny, Hari Sundaram, and Varun Chandrasekaran. Training robust ensembles requires rethinking lipschitz continuity. In *International Conference on Learning Representations (ICLR)*, 2025.

[Essameldin *et al.*, 2020] Aliaa Essameldin, Mohamed Nurulhoque, and Khaled A Harras. More than the sum of its things: Resource sharing across iots at the edge. In *ACM/IEEE SEC*, 2020.

[Filters'Importance, 2016] Determine Filters'Importance. Pruning filters for efficient convnets. 2016.

[Fontanesi *et al.*, 2025] Gianluca Fontanesi, Flor Ortíz, Eva Lagunas, Luis Manuel Garcés-Socarrás, Victor Monzon Baeza, Miguel Ángel Vázquez, Juan Andrés Vásquez-Peralvo, Mario Minardi, Ha Nguyen Vu, Puneeth Jubba Honnaiah, et al. Artificial intelligence for satellite communication: A survey. *IEEE Communications Surveys & Tutorials*, 2025.

[Frankle and Carlin, 2019] Jonathan Frankle and Michael Carlin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019.

[Gao *et al.*, 2025] Jingxiang Gao, Hend K Gedawy, Eduardo Feo-Flushing, and Khaled A Harras. A deployable privacy-preserving thermal-based obstacle detection system for indoor navigation. In *IEEE ICC*, 2025.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Gui *et al.*, 2019] Shupeng Gui, Haotao Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, and Ji Liu. Model compression with adversarial robustness: A unified optimization framework. In *Advances in Neural Information Processing Systems*, 2019.

[Han *et al.*, 2015] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems*, 2015.

[Jung and Song, 2025] Yoojin Jung and ByungCheol Song. Two is better than one: Efficient ensemble defense for robust and compact models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9696–9705, 2025.

[Kariyappa and Qureshi, 2019] Sanjay Kariyappa and Moinuddin K Qureshi. Improving adversarial robustness of ensembles with diversity training. In *arXiv preprint*, 2019.

[Kwon *et al.*, 2021] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pages 5905–5914. PMLR, 2021.

[Lee *et al.*, 2022] Byung-Kwan Lee, Junho Kim, and Yong Man Ro. Masking adversarial damage: Finding adversarial saliency for robust and sparse network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15126–15136, 2022.

[Li *et al.*, 2023] Chenhao Li, Qiang Qiu, et al. Learning adversarially robust sparse networks via weight reparameterization. In *AAAI Conference on Artificial Intelligence*, 2023.

[Liu *et al.*, 2017] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[Madry *et al.*, 2018] 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.

[Mostafa *et al.*, 2024] Sherif Mostafa, Moustafa Youssef, and Khaled A Harras. Accurate and ubiquitous floor identification at the edge using a single cell tower. In *IEEE/ACM SEC*, 2024.

[Mostafa *et al.*, 2026] Sherif Mostafa, Khaled A Harras, and Moustafa Youssef. Human-as-a-sensor: Toward autonomous multimodal sensing for self-determined technology engagement. *IEEE Pervasive Computing*, 2026.

[Pang *et al.*, 2019] Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, 2019.

[Piras *et al.*, 2025] Giorgio Piras, Maura Pintor, Ambra Demontis, Battista Biggio, Giorgio Giacinto, and Fabio Roli. Adversarial pruning: A survey and benchmark of pruning methods for adversarial robustness. *Pattern Recognition*, page 111788, 2025.

[Saeed *et al.*, 2017] Ahmed Saeed, Ahmed Abdelkader, Mouhyemen Khan, Azin Neishaboori, Khaled A Harras, and Amr Mohamed. Argus: realistic target coverage by drones. In *ACM/IEEE IPSN*, 2017.

[Saeed *et al.*, 2019] Ahmed Saeed, Ahmed Abdelkader, Mouhyemen Khan, Azin Neishaboori, Khaled A Harras, and Amr Mohamed. On realistic target coverage by autonomous drones. *ACM Transactions on Sensor Networks (TOSN)*, 15(3):1–33, 2019.

[Saxena *et al.*, 2024] Divya Saxena, Jiannong Cao, Jiahao Xu, and Tarun Kulshrestha. Rg-gan: Dynamic regenerative pruning for data-efficient generative adversarial networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4704–4712, 2024.

[Sehwag *et al.*, 2020] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. In *Advances in Neural Information Processing Systems*, 2020.

[Suresh *et al.*, 2025] Janani Suresh, Nancy Nayak, and Sheetal Kalyani. First line of defense: A robust first layer mitigates adversarial attacks. In *AAAI Conference on Artificial Intelligence (AAAI-25)*, pages 7176–7184, 2025.

[Tsuzuku *et al.*, 2018] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[Wang *et al.*, 2019] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2019.

[Waseda *et al.*, 2025] Futa Waseda, Ching-Chun Chang, and Isao Echizen. Rethinking invariance regularization in adversarial training to improve robustness-accuracy tradeoff. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2402.14648.

[Yang *et al.*, 2020] Huanrui Yang, Jingyang Zhang, Hongliang Dong, et al. Dverge: Diversifying vulnerabilities for enhanced robust generation of ensembles. In *NeurIPS*, 2020.

[Ye *et al.*, 2019] Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[Zha *et al.*, 2025] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025.

[Zhang *et al.*, 2018] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018.

[Zhang *et al.*, 2019] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

[Zhang *et al.*, 2023] Lei Zhang, Zhibo Wang, Xiaowei Dong, Yunhe Feng, Xiaoyi Pang, Zhifei Zhang, and Kui Ren. Towards fairness-aware adversarial network pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5168–5177, 2023.

[Zhang *et al.*, 2024] Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. Beyond interpretability: The gains of feature monosemanticity on model robustness. *arXiv preprint arXiv:2410.21331*, 2024.

[Zhao and Wressnegger, 2024] Qi Zhao and Christian Wressnegger. Holistic adversarially robust pruning. *arXiv preprint arXiv:2412.14714*, 2024.

[Zhong *et al.*, 2025] Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20180–20192, 2025.

[Zhou *et al.*, 2025] Nuoyan Zhou, Dawei Zhou, Decheng Liu, Nannan Wang, and Xinbo Gao. Mitigating feature gap for adversarial robustness by feature disentanglement. In *AAAI Conference on Artificial Intelligence (AAAI-25)*, pages 10825–10833, 2025.