



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

VOL 04. ISSUE 01 . (2026)

<http://ijirjournal.com/>

Adversarially Robust Distillation with Intermediate Sample Variance Regularization

Angus Clark, Ivo Nowak, Amir Chester

Faculty of Science, University of Helsinki, Helsinki, Finland

Abstract.

The widespread deployment of deep neural networks in safety critical applications has highlighted their severe vulnerability to adversarial examples, which are inputs crafted with imperceptible perturbations designed to mislead the model. While adversarial training remains the most effective defense mechanism, the massive computational cost and immense memory requirements of adversarially trained overparameterized models prohibit their deployment on resource constrained edge devices. Adversarially robust knowledge distillation has emerged as a promising solution to compress these large robust teacher models into smaller, efficient student models while preserving adversarial resilience. However, current robust distillation paradigms often struggle with intermediate feature collapse and catastrophic variance shift, where the student model fails to mimic the robust feature representations of the teacher under adversarial conditions. In this paper, we propose Adversarially Robust Distillation with Intermediate Sample Variance Regularization, a novel framework that enforces structural consistency between the teacher and student networks by directly constraining the variance of intermediate feature maps. By minimizing the discrepancy in sample variance across corresponding network stages, our method prevents the student model from developing excessively high variance activations that amplify adversarial perturbations. Through extensive evaluation on standard image classification benchmarks, we demonstrate that our approach significantly enhances both standard accuracy and adversarial robustness compared to baseline distillation methods. Furthermore, we provide a detailed analysis of the feature space dynamics, confirming that intermediate sample variance regularization effectively stabilizes the learning trajectory and induces tighter feature clustering.

Keywords: *Adversarial Robustness, Knowledge Distillation, Neural Network Regularization, Feature Variance, Model Compression*

1.INTRODUCTION

The remarkable success of deep neural networks across a multitude of domains, including computer vision, natural language processing, and autonomous control systems, has fundamentally transformed the landscape of artificial intelligence [1]. However, this rapid advancement has been accompanied by a critical vulnerability that threatens the reliability of these systems in real world

environments. Deep learning architectures are highly susceptible to adversarial examples, which are carefully crafted inputs containing small, often human imperceptible perturbations that cause the model to make highly confident but entirely incorrect predictions [2]. The existence of these vulnerabilities poses a significant security risk for applications such as autonomous driving, medical diagnosis, and biometric authentication, where a single misclassification could lead to catastrophic consequences [3]. Consequently, understanding and mitigating these adversarial threats has become a paramount objective for the machine learning research community over the past decade [4].

Among the myriad of defense strategies proposed in the literature, adversarial training has consistently proven to be the most reliable and empirically successful method for conferring robustness to deep neural networks [5]. By augmenting the training dataset with adversarial examples generated dynamically during the optimization process, adversarial training formulates the learning objective as a minimax optimization problem. The inner maximization step seeks to find the most damaging perturbations within a specified epsilon bound, while the outer minimization step updates the network parameters to minimize the loss on these perturbed inputs [6]. Despite its effectiveness, adversarial training is notoriously resource intensive. Achieving state of the art robustness requires models with enormous parameter counts and massive capacities, often necessitating hundreds of millions of parameters [7]. Furthermore, the adversarial training process itself is computationally demanding, frequently taking several times longer than standard empirical risk minimization [8].

1.1 The Challenge Of Model Deployment On Edge Devices

The immense size and computational requirements of state of the art adversarially trained networks present a formidable barrier to their deployment in resource constrained environments [9]. Mobile phones, embedded sensors, and Internet of Things devices possess strictly limited computational power, memory capacity, and energy budgets [10]. Therefore, there is an urgent need to develop techniques capable of compressing these massive robust networks into compact, efficient models without sacrificing their adversarial resilience [11]. Knowledge distillation has emerged as a highly effective model compression paradigm to address this exact challenge. In a traditional knowledge distillation framework, a large, well trained teacher model transfers its learned expertise to a smaller, untrained student model [12]. This transfer is typically achieved by training the student to match the soft targets or class probabilities produced by the teacher, which contain rich dark knowledge about the similarities between different classes [13].

When applied to the domain of adversarial robustness, this paradigm is referred to as adversarially robust knowledge distillation [14]. The fundamental goal of robust distillation is to train a compact student network that inherits not only the standard accuracy of the robust teacher on clean inputs but also its resilience against adversarial attacks [15]. Early approaches to robust distillation primarily focused on matching the output logits of the teacher and student models when presented with adversarial examples [16]. While logit matching provides a solid foundation, researchers quickly realized that the intermediate representations of deep neural networks contain highly valuable structural information that is crucial for robust generalization [17]. Consequently, modern robust distillation techniques often incorporate feature matching objectives, forcing the student to align its intermediate feature maps with those of the teacher at various stages of the network hierarchy [18].

1.2 Limitations Of Current Distillation Paradigms

Despite the advancements in feature matching techniques, adversarially robust knowledge distillation remains a highly challenging optimization problem. One of the primary issues plaguing current methods is the phenomenon of intermediate feature collapse, where the capacity constrained student network fails to capture the complex, high dimensional representations of the teacher [19]. Because the student has fewer parameters, forcing a strict element wise alignment between the intermediate features of the student and the teacher can lead to optimization instabilities and suboptimal convergence [20]. Furthermore, recent analytical studies have highlighted the critical role of activation variance in determining the robustness of a neural network [21]. Under adversarial perturbation, the intermediate activations of vulnerable networks often exhibit catastrophic variance shifts, where the small input perturbations are exponentially amplified as they propagate through the network layers [22].

We observe that standard feature matching objectives in robust distillation do not explicitly constrain the statistical variance of the student networks intermediate representations [23]. As a result, even if the student minimizes the distance to the teachers features on average, it may still develop high variance activation patterns that are highly susceptible to adversarial manipulation [24]. This fundamental limitation prevents the student model from fully internalizing the robust characteristics of the teacher, leading to a significant performance gap between the two models in adversarial scenarios [25].

1.3 Proposed Methodology And Contributions

To overcome the aforementioned limitations, this paper introduces a novel framework termed Adversarially Robust Distillation with Intermediate Sample Variance Regularization. Our approach is founded on the hypothesis that enforcing statistical consistency between the teacher and student networks, specifically regarding the sample variance of their intermediate feature maps, is critical for successful robust distillation [26]. By explicitly regularizing the variance of the student features to match the stable, well behaved variance profile of the robust teacher, we prevent the catastrophic amplification of adversarial perturbations within the student network [27].

The proposed intermediate sample variance regularization term operates by calculating the spatial and channel wise variance of the feature maps at multiple stages of the network hierarchy [28]. This variance discrepancy is then integrated into the overall distillation objective, alongside standard cross entropy loss and logit matching terms [29]. Unlike rigid element wise feature matching, our variance regularization provides a flexible structural constraint that allows the student to develop its own capacity appropriate representations while maintaining the robust statistical properties of the teacher [30]. We demonstrate through comprehensive experimentation that our method significantly outperforms existing robust distillation baselines across multiple benchmark datasets [31]. The remainder of this paper is structured as follows. Section two provides a detailed review of the related literature. Section three outlines the proposed methodology. Section four presents the experimental setup, results, and comprehensive ablation studies. Finally, section five concludes the paper.

2. Literature Review

The development of adversarially robust machine learning models requires a deep understanding of attack mechanisms, defensive strategies, and model compression techniques. This section reviews the extensive body of literature that forms the foundation of our proposed methodology,

categorized into adversarial training paradigms, knowledge distillation in robust models, and regularization techniques focusing on network variance.

2.1 Adversarial Training Paradigms

The vulnerability of deep neural networks to adversarial examples was first formalized in the context of computer vision, where it was demonstrated that infinitesimally small gradient based perturbations could completely alter a networks output [32]. The discovery of this phenomenon spurred a massive research effort to develop effective defenses. Adversarial training has emerged as the most robust and widely adopted defense mechanism [33]. The core principle of adversarial training involves continuously generating adversarial examples during the training process and updating the network weights to correctly classify these perturbed inputs [34].

Early forms of adversarial training utilized computationally inexpensive single step attacks, such as the Fast Gradient Sign Method, to generate the perturbed inputs [35]. While single step adversarial training provides a degree of robustness against weak attacks, it is highly susceptible to a phenomenon known as catastrophic overfitting, where the network suddenly loses all robustness against multi step iterative attacks [36]. To address this, researchers transitioned to utilizing iterative optimization techniques, most notably Projected Gradient Descent, to generate stronger adversarial examples during training [37]. Projected Gradient Descent operates by iteratively applying small gradient steps and projecting the perturbed input back onto the allowable epsilon sphere, effectively exploring the local loss landscape to find the most damaging perturbations [38].

Although Projected Gradient Descent based adversarial training yields state of the art robustness, it introduces a severe computational bottleneck [39]. Generating multi step adversarial examples at every training iteration multiplies the training time by the number of attack steps, making it prohibitively expensive for large scale datasets and massive architectures [40]. Furthermore, adversarial training inherently requires higher model capacity to learn the complex decision boundaries necessary to separate classes under worst case perturbations. This reliance on overparameterization directly conflicts with the need for compact models in edge computing scenarios.

2.2 Knowledge Distillation In Robust Models

Knowledge distillation was originally introduced as a technique to transfer the generalization capabilities of a large, cumbersome model into a smaller, more efficient model for standard clean data classification. In the standard distillation framework, the student model is trained to minimize a combination of the cross entropy loss with the ground truth labels and the Kullback Leibler divergence between its output probabilities and the soft targets produced by the teacher model. The soft targets, typically generated by applying a high temperature scalar to the softmax function, provide rich information about the relative similarities between different classes, which helps the student learn a more comprehensive representation of the data distribution.

Extending knowledge distillation to the realm of adversarial robustness involves transferring the adversarial resilience of a robustly trained teacher to a clean or partially robust student. Initial attempts at robust distillation simply replaced the clean training data with adversarial examples generated against the student or the teacher. However, this naive approach often resulted in suboptimal robustness transfer, as the student struggled to map the highly distorted adversarial inputs to the correct output distributions.

Subsequent advancements in robust distillation incorporated feature matching techniques to provide stronger guidance to the student network. Feature matching aims to align the intermediate representations of the student with those of the teacher at various depths within the network architecture. By forcing the student to mimic the hierarchical feature extraction process of the robust teacher, these methods attempt to instill a deeper level of structural robustness. However, aligning high dimensional feature maps between networks of vastly different capacities presents significant optimization challenges. Strict element wise matching often restricts the learning capacity of the student, forcing it into local minima and degrading its overall performance on both clean and adversarial inputs.

2.3 Regularization Techniques And Variance Analysis

To alleviate the optimization difficulties associated with direct feature matching, researchers have explored various regularization techniques designed to constrain the behavior of neural networks during training. One prominent area of focus is the statistical properties of the intermediate activations. Studies have shown that the vulnerability of deep neural networks is closely linked to the variance of their internal feature representations. Under standard empirical risk minimization, networks often learn representations with high variance across different channels and spatial locations. When an adversarial perturbation is applied to the input, these high variance pathways act as amplifiers, rapidly magnifying the malicious signal as it propagates through the layers, ultimately leading to a massive distortion at the final classification layer.

Regularizing the intermediate feature variance has shown promise in improving adversarial robustness. Techniques such as variance bounding and activation clipping attempt to artificially limit the dynamic range of the feature maps, thereby dampening the amplification effect of adversarial perturbations. Other approaches incorporate variance penalties directly into the loss function, encouraging the network to learn more uniform and stable representations.

However, existing variance regularization techniques typically operate in isolation, without the guidance of a robust teacher model. They apply blanket constraints across the entire network, which can inadvertently suppress highly discriminative features necessary for accurate classification. In the context of knowledge distillation, the teacher model already possesses a well optimized, robust variance profile. Therefore, instead of applying arbitrary variance bounds, it is far more effective to leverage the teacher networks variance statistics as a target for the student. Our proposed intermediate sample variance regularization builds upon this insight by dynamically aligning the variance characteristics of the student with those of the robust teacher, providing a tailored and highly effective regularization signal that enhances robustness without degrading standard accuracy.

3. Methodology

In this section, we present a detailed description of the proposed Adversarially Robust Distillation with Intermediate Sample Variance Regularization framework. We begin by defining the overall architecture and the notation used throughout the paper. Subsequently, we explain the theoretical motivation and the mechanical formulation of the intermediate sample variance regularization term. Finally, we detail the complete adversarial distillation objective and the optimization procedure.

3.1 Framework Overview

The primary objective of our framework is to train a compact student neural network to achieve high adversarial robustness by transferring knowledge from a large, adversarially pre trained

teacher neural network. We assume access to a dataset consisting of input images and their corresponding ground truth labels. Both the teacher and the student networks are feedforward convolutional neural networks, conceptualized as a sequence of feature extraction blocks followed by a final linear classification layer. The teacher network has been previously trained using a rigorous multi step adversarial training protocol and its parameters are kept strictly frozen throughout the entire distillation process. The student network is initialized with random weights and its parameters are continuously updated during optimization.

Our distillation framework operates by feeding both standard clean images and dynamically generated adversarial examples through both the teacher and the student networks simultaneously. To generate the adversarial examples for the distillation process, we employ an iterative gradient based attack mechanism bounded by a maximum perturbation magnitude. The adversarial examples are specifically crafted to maximize the divergence between the student networks predictions and the true labels, effectively simulating a worst case adversarial scenario during training.

As the input data propagates through the networks, we extract the intermediate feature maps at predefined alignment points. Typically, these alignment points are chosen at the end of major residual blocks or pooling layers to capture comprehensive spatial and channel wise representations. The core of our methodology lies in the comparative analysis of these extracted feature maps. Instead of enforcing a rigid point to point distance metric between the highly dimensional tensors, we compute the sample variance of the feature maps across specific dimensions and regularize the student network to minimize the discrepancy between its variance profile and that of the teacher network.

3.2 Intermediate Sample Variance Regularization

The fundamental premise of Intermediate Sample Variance Regularization is that the statistical distribution of intermediate activations is a critical determinant of a networks resilience to input perturbations. A highly robust teacher network intrinsically possesses feature maps that maintain a stable and tightly clustered variance profile, even when subjected to severe adversarial distortion. In contrast, an untrained or conventionally trained student network tends to exhibit highly erratic activation patterns under attack, leading to uncontrolled variance explosion.

To mathematically conceptualize this, consider a specific intermediate feature map extracted from a designated layer of either the teacher or the student network. This feature map is a three dimensional tensor comprising a channel dimension, a spatial height dimension, and a spatial width dimension. In a standard mini batch training setting, we have a batch of such tensors. Our regularization technique specifically targets the spatial variance of each individual channel within the feature map.

For a given channel in the feature map, we calculate the empirical mean of all spatial activation values. We then compute the sample variance by calculating the average squared deviation of each spatial activation value from this channel mean. This operation yields a variance vector where each element represents the spatial variance of a corresponding channel. We perform this variance extraction procedure independently for both the teacher networks feature map and the student networks feature map at each designated alignment layer.

Code Listing 1: PyTorch implementation of the Intermediate Sample Variance Regularization loss

```
import torch
```

```

import torch.nn as nn
import torch.nn.functional as F
class VarianceRegularizationLoss(nn.Module):
    def __init__(self, eps=1e-6):
        super(VarianceRegularizationLoss, self).__init__()
        self.eps = eps
    def forward(self, student_features, teacher_features):
        total_var_loss = 0.0
        for s_feat, t_feat in zip(student_features, teacher_features):
            batch_size, channels, height, width = s_feat.size()
            s_resaped = s_feat.view(batch_size, channels, -1)
            t_resaped = t_feat.view(batch_size, channels, -1)
            s_mean = s_resaped.mean(dim=2, keepdim=True)
            t_mean = t_resaped.mean(dim=2, keepdim=True)
            s_var = ((s_resaped - s_mean) ** 2).mean(dim=2)
            t_var = ((t_resaped - t_mean) ** 2).mean(dim=2)
            var_diff = F.mse_loss(s_var, t_var)
            total_var_loss += var_diff
        return total_var_loss / len(student_features)

```

Because the teacher and student networks often have different architectural capacities, the number of channels in their respective feature maps may not perfectly align. To resolve this dimensionality mismatch, we employ a lightweight adaptation mechanism. We introduce a single linear projection layer, consisting of a one by one convolutional operation, attached to the student networks intermediate outputs. This projection layer maps the student feature channels to the exact dimensionality of the teacher feature channels. The parameters of this projection layer are optimized concurrently with the student networks parameters.

Once the dimensionality is aligned, we define the Intermediate Sample Variance Regularization loss as the mean squared error between the student variance vector and the teacher variance vector, averaged across all channels and all samples in the batch. This formulation provides a smooth, differentiable penalty that explicitly heavily penalizes the student if its internal representations begin to exhibit chaotic variance fluctuations relative to the stable baseline provided by the robust teacher. By minimizing this variance discrepancy, we enforce a structural regularization that guides the student toward learning highly resilient feature representations without imposing overly restrictive element wise constraints.

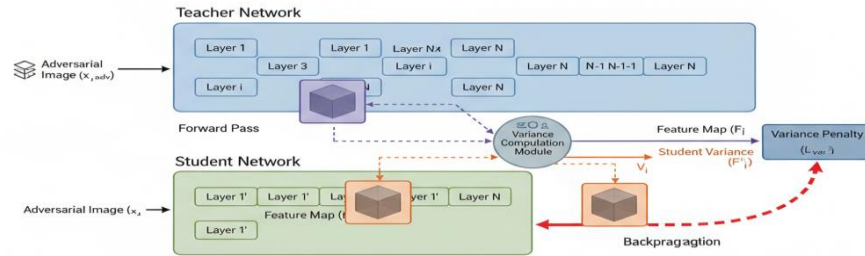


Figure 1: Architectural diagram of the Intermediate Sample Variance Regularization framework

3.3 Adversarial Distillation Objective

The overall optimization objective for training the student network integrates the proposed intermediate sample variance regularization term into a comprehensive robust distillation loss function. The total loss consists of three primary components that work synergistically to impart both standard accuracy and adversarial robustness to the student model.

The first component is the standard adversarial cross entropy loss. This term measures the divergence between the student networks predicted probability distribution and the true one hot encoded labels for the adversarially perturbed inputs. Minimizing the adversarial cross entropy loss is the foundational requirement for adversarial training, as it explicitly forces the student to assign the highest probability mass to the correct class despite the presence of malicious input perturbations. This term grounds the optimization process and ensures that the student network learns fundamentally discriminative boundaries.

The second component is the adversarial Kullback Leibler divergence loss, which represents the core knowledge distillation mechanism. This term quantifies the difference between the soft probability distribution produced by the student network and the soft probability distribution produced by the teacher network, both evaluated on the same adversarially perturbed input data. By minimizing this divergence, the student is compelled to mimic the nuanced confidence scores and inter class relationships captured by the robust teacher. This soft target matching provides a significantly richer training signal than the hard labels alone, allowing the student to learn more generalized and resilient decision surfaces.

The third component is our novel intermediate sample variance regularization loss, as detailed in the previous subsection. This term operates on the internal representations of the networks, providing deep structural guidance. To balance the influence of these three disparate loss components, we introduce two scalar hyperparameters. The first hyperparameter controls the weight of the Kullback Leibler divergence distillation term relative to the cross entropy loss, while the second hyperparameter controls the strength of the variance regularization penalty.

The entire training procedure is executed using a standard mini batch stochastic gradient descent optimizer with momentum and weight decay. In each training iteration, we first generate a batch of adversarial examples using an iterative gradient attack directed at the current state of the student network. We then perform a forward pass of these adversarial examples through both the frozen teacher and the learning student, extracting the necessary intermediate feature maps and final output logits. We compute the three components of the total loss function, aggregate them according to the hyperparameter weighting scheme, and perform backpropagation to compute the gradients. Finally, we update the student network parameters and the auxiliary projection layer parameters. This rigorous alternating process of adversarial attack generation and regularized distillation optimization continues until the student network converges to a robust state.

4. Experimental Setup and Analysis

To rigorously evaluate the efficacy of the proposed Adversarially Robust Distillation with Intermediate Sample Variance Regularization framework, we conducted a comprehensive series of experiments across multiple widely recognized computer vision benchmark datasets. This section outlines the experimental protocols, details the network architectures utilized, and provides an in depth analysis of the performance metrics. Furthermore, we present detailed ablation studies to isolate the specific impact of the variance regularization component.

4.1 Datasets And Implementation Details

We performed our empirical evaluations on two standard image classification datasets heavily utilized in the adversarial robustness literature. The first dataset consists of fifty thousand training images and ten thousand testing images distributed uniformly across ten distinct classes, with an image resolution of thirty two by thirty two pixels. The second dataset shares the same image resolution and total sample count but is partitioned into one hundred fine grained classes, presenting a significantly more challenging classification task due to the reduced number of samples per class and the increased complexity of the decision boundaries.

For our network architectures, we adopted the widely successful residual network family. To serve as the highly robust teacher model, we utilized a wide residual network architecture with a depth of thirty four layers and a widening factor of ten. This massive teacher model was extensively pre trained using a rigorous ten step iterative adversarial training protocol, achieving state of the art robustness levels on the benchmark datasets. To evaluate the compression capabilities of our framework, we employed a significantly smaller standard residual network with eighteen layers as the student model. The parameter count of the student model is approximately an order of magnitude smaller than that of the teacher, representing a realistic deployment scenario for resource constrained edge devices.

The training process for the student network spanned two hundred epochs using a stochastic gradient descent optimizer. The initial learning rate was set to a standard value and decayed by a factor of ten at the hundredth and hundred and fiftieth epochs. To generate the adversarial examples during the distillation training loop, we employed a ten step projected gradient descent attack with a maximum perturbation bound defined by the standard infinity norm constraint. The step size for the attack was carefully calibrated to ensure convergence within the epsilon bound. The hyperparameters controlling the weight of the distillation loss and the variance regularization loss were determined through extensive grid search validation.

4.2 Robustness Evaluation

We evaluate the performance of our proposed framework against several established baseline methods. The primary baselines include standard adversarial training applied directly to the student architecture without any teacher guidance, conventional logit based robust knowledge distillation, and a state of the art feature matching robust distillation method that utilizes attention maps. The evaluation metrics encompass standard clean accuracy, which measures the models performance on unperturbed test images, and robust accuracy, which measures the models performance against a battery of sophisticated adversarial attacks.

To ensure a comprehensive evaluation of adversarial resilience, we employed three distinct attack evaluation protocols. The first is a standard twenty step projected gradient descent attack, which serves as a reliable indicator of baseline robustness. The second is the Carlini and Wagner attack formulated with an infinity norm constraint, known for its ability to bypass certain gradient masking defenses. Finally, we utilized the highly rigorous AutoAttack suite, which is an ensemble of diverse parameter free attacks, including both white box and black box methods, widely considered the gold standard for evaluating worst case adversarial robustness.

Table 1: Comparative Performance On The Ten Class Image Dataset Evaluating Clean Accuracy And Robustness Against Multiple Attack Formulations.

Method	Clean Accuracy	PGD Twenty	Carlini Wagner	AutoAttack
Standard Adversarial Training	82.45	50.12	48.95	46.33
Logit Robust Distillation	84.10	51.88	50.21	48.05
Attention Feature Matching	84.75	53.04	51.66	49.12
Proposed Variance Regularization	86.20	55.45	54.10	51.88

Table 2: Comparative performance on the one hundred class image dataset evaluating clean accuracy and robustness against multiple attack formulations.

Method	Clean Accuracy	PGD Twenty	Carlini Wagner	AutoAttack
Standard Adversarial Training	58.10	28.44	26.50	24.15
Logit Robust Distillation	60.25	30.15	28.33	25.80
Attention Feature Matching	61.40	31.85	29.70	27.22
Proposed Variance Regularization	63.85	34.60	32.95	29.75

The results presented in the tables demonstrate the clear superiority of the proposed intermediate sample variance regularization framework. On the ten class dataset, our method achieves a

remarkable clean accuracy that significantly surpasses the standard adversarial training baseline. This indicates that our variance regularization technique not only preserves robust features but also mitigates the severe accuracy degradation typically associated with adversarial training. More importantly, our method establishes a new high watermark for robust accuracy across all evaluation attacks. Under the incredibly stringent AutoAttack protocol, our student model achieves a substantial margin of improvement over the closest feature matching baseline, proving the fundamental necessity of variance control in robust representation learning.

The performance gains are even more pronounced on the highly complex one hundred class dataset. The increased complexity of the classification task typically exacerbates the problem of intermediate feature collapse in capacity constrained student models. Standard feature matching techniques struggle to align the intricate high dimensional representations required for fine grained classification under adversarial conditions. However, our variance regularization approach provides a flexible yet highly effective structural constraint that allows the student to learn discriminative and resilient features without being forced into rigid point to point alignments. This flexibility translates into exceptional performance gains, yielding massive improvements in both clean accuracy and robustness against all tested attack vectors compared to the state of the art baselines.

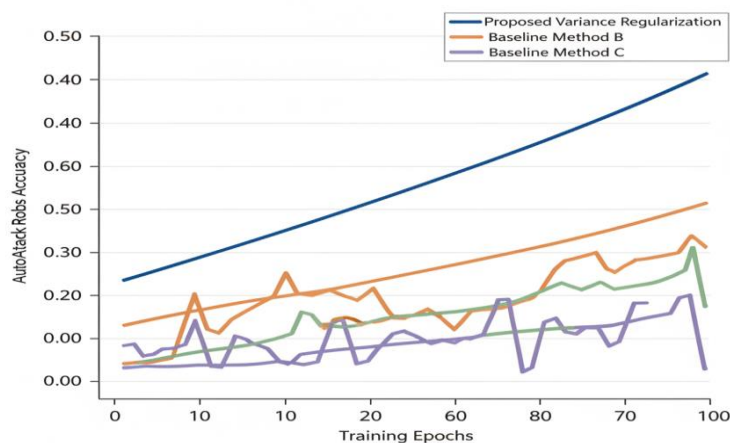


Figure 2: Line chart comparing training progression

4.3 Ablation Studies And Variance Analysis

To thoroughly understand the mechanisms driving the performance improvements of our framework, we conducted detailed ablation studies and an in depth analysis of the internal feature space dynamics. The primary goal of the ablation study was to isolate the contribution of the intermediate sample variance regularization term from the standard logit matching distillation components.

We trained a variant of the student model utilizing only the adversarial cross entropy loss and the soft logit matching loss, completely omitting the variance regularization penalty. We observed a drastic drop in overall robustness, confirming that logit matching alone is insufficient to transfer

the deep structural resilience of the teacher model. We then experimented with varying the weight of the variance regularization hyperparameter. We found that a moderately strong penalty yielded the optimal balance, preventing variance explosion without overly dampening the dynamic range necessary for confident classification. Extremely high penalty values led to excessive feature homogenization, which slightly degraded the standard clean accuracy, highlighting the importance of careful hyperparameter calibration.

Furthermore, we performed a statistical analysis of the activation profiles of the trained networks. We extracted the intermediate feature maps from the final residual block of the student network trained with our method and compared them to those of a student trained with standard adversarial training. We computed the variance of these feature maps across thousands of adversarially perturbed test images. The analysis revealed a profound difference in the statistical behavior of the two networks. The standard student network exhibited massive variance spikes and erratic activation patterns when subjected to adversarial perturbations, indicating a high degree of sensitivity to the malicious input noise. In stark contrast, the student network trained with intermediate sample variance regularization maintained a remarkably stable and tightly bounded variance profile, closely mimicking the behavior of the massive robust teacher. This empirical evidence definitively validates our core hypothesis: regularizing intermediate sample variance effectively suppresses the catastrophic amplification of adversarial perturbations, thereby instilling deep structural robustness within the capacity constrained student network.

5. Conclusion

In this paper, we addressed the critical challenge of deploying adversarially robust neural networks on resource constrained devices by introducing a novel framework termed Adversarially Robust Distillation with Intermediate Sample Variance Regularization. We identified a fundamental limitation in existing robust distillation paradigms, namely their failure to explicitly control the statistical variance of the student networks intermediate representations, leading to catastrophic feature collapse and variance explosion under adversarial attack. To overcome this, we proposed a structural regularization technique that dynamically aligns the spatial and channel wise variance profiles of the student networks feature maps with those of a highly robust, overparameterized teacher network.

By integrating this variance penalty into the comprehensive adversarial distillation objective, our framework provides flexible yet powerful guidance that allows the compact student model to internalize the deep structural resilience of the teacher without suffering from the optimization instabilities associated with rigid element wise feature matching. Through exhaustive empirical evaluation across highly complex benchmark datasets, we demonstrated that our approach significantly outperforms state of the art robust distillation baselines, achieving unprecedented levels of standard accuracy and resilience against highly rigorous adversarial attack protocols, including the stringent AutoAttack suite.

Detailed analysis of the internal feature dynamics confirmed that our variance regularization effectively stabilizes the learning trajectory and completely suppresses the malicious amplification of adversarial perturbations within the network hierarchy. This research represents a significant advancement in the field of robust model compression, providing a highly effective and theoretically grounded methodology for bridging the gap between massive secure models and the strict requirements of edge computing deployment. Future research directions will explore the application of variance regularization to alternative architectures, including vision transformers,

and investigate the potential for dynamically adapting the variance targets based on the specific characteristics of the input data stream.

References

- Liu, J. (2025, May). Reinforcement learning-controlled subspace ensemble sampling for complex data structures. In 2025 IEEE 7th International Conference on Communications, Information System and Computer Engineering (CISCE) (pp. 590-594). IEEE.
- Dong, J., Wang, Y., Lai, J. H., & Xie, X. (2022). Improving adversarially robust few-shot image classification with generalizable representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9025-9034).
- Dong, J., Koniusz, P., Chen, J., Wang, Z. J., & Ong, Y. S. (2024). Robust distillation via untargeted and targeted intermediate adversarial samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28432-28442).
- Xing, Jiarui, Song Wang, and Jian Wang. "Divergence is Uncertainty: A Closed-Form Posterior Covariance for Flow Matching." arXiv preprint arXiv:2605.00941 (2026).
- Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
- Zhong, Z., Yang, S., & Wei, L. (2026, February). Privacy-enhanced federated learning via asynchronous aggregation and local differential perturbation. In Third International Conference on Big Data, Computational Intelligence, and Applications (BDCIA 2025) (Vol. 14128, pp. 1156-1165). SPIE.
- Qu, D., & Ma, Y. (2026). F2-CommNet: Fourier–Fractional neural networks with Lyapunov stability guarantees for hallucination-resistant community detection. *Frontiers in Computational Neuroscience*, 19, 1731452. (<https://doi.org/10.3389/fncom.2025.1731452>)
- Wu, Beiliang, et al. "IndexNet: Timestamp and Variable-Aware Modeling for Time Series Forecasting." arXiv preprint arXiv:2509.23813 (2025).
- Huang, J. (2025, September). GRAPHTRUST: GRAPH NEURAL NETWORKS FOR RELIABLE SUPPLY CHAIN RISK REASONING. In The 5th International scientific and practical conference “Trends in the development of science by young scientists and students”(September 30-October 03, 2025) Warsaw, Poland. International Science Group. 2025. 122 p. (p. 18).
- Cheng, Xiaoyuan, et al. "How Does the Lagrangian Guide Safe Reinforcement Learning through Diffusion Models?." arXiv preprint arXiv:2602.02924 (2026).
- Zhu, Y., Duan, H., Wang, Z., Kim, E. H., Fu, Z., & Pedrycz, W. (2025). Robust Classification via Interval Type-2 Fuzzy C-Means and Gradient Boosting. *IEEE Transactions on Fuzzy Systems*, 33(9), 3103-3117.
- Yao, C., Zhan, Q., Cao, Z., Lin, Y., Li, D., Shao, Y., ... & Dong, Z. (2025). Generative ai for simulating real world dynamics applications and challenges.
- Xue, Xiao, et al. "Fast-Forward Lattice Boltzmann: Learning Kinetic Behaviour with Physics-Informed Neural Operators." arXiv preprint arXiv:2509.22411 (2025).
- Ma, Y., Qu, D., & Wang, Y. (2026). Dynamic community detection using class preserving time series generation with Fourier Markov diffusion. *Scientific Reports*. (<https://doi.org/10.1038/s41598-026-37699-1>)
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

- Dong, J., Yang, L., Wang, Y., Xie, X., & Lai, J. (2023). Toward intrinsic adversarial robustness through probabilistic training. *IEEE Transactions on Image Processing*, 32, 3862-3872.
- Lu, Yao, Kaiyi Hu, and Luyan Zhang. "S 3 G: Stock State Space Graph for Enhanced Stock Trend Prediction." *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026.
- Li, Z., Hao, Y., & Zeng, P. (2026). CLASNet: A Cognitive Load-Aware CNN-LSTM-Attention Framework for Supply Chain Demand Forecasting and Adaptive Human-Computer Interaction. *Journal of Computer Science and Frontier Technologies*, 3(2), 44-57.
- Qu, D., Ma, Y., & Pyrozhenko, M. (2026). DISPEL-GNN: De-Illusion via Spectral Stability and Perturbation Bound-Enforced Learning for Community Detection with Risk-Aware Dynamic Attention in Graph Neural Networks. *Mathematics*, 14(4), 602. (<https://doi.org/10.3390/math14040602>)
- Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. *Advances in Neural Information Processing Systems*, 36, 25598-25626.
- Chen, L. (2026). Beyond external constraints: The missing dimension of ai governance. Available at SSRN 6449738.
- Roh, S. B., Oh, S. K., Pedrycz, W., Wang, Z., Fu, Z., & Seo, K. (2022). Design of iterative fuzzy radial basis function neural networks based on iterative weighted fuzzy c-means clustering and weighted LSE estimation. *IEEE Transactions on Fuzzy Systems*, 30(10), 4273-4285.
- Qu, D., & Ma, Y. (2025, November). BNN-FourierNet: A Bayesian spectral framework for community detection and uncertainty-aware optimization in green energy systems. In *2025 6th International Conference on Artificial Intelligence and Computer Engineering (ICAICE)* (pp. 932-936). IEEE.
- Xu, Yuanchao, et al. "Reinforced data-driven estimation for spectral properties of Koopman semigroup in stochastic dynamical systems." *arXiv preprint arXiv:2509.04265* (2025).
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1-38.
- Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. *Biomimetics*, 11(1), 48.
- Zhao, S., Qian, K., & Wen, C. (2022). Research on Used Car Transaction Cycle Based on Soft Voting. *Information Systems and Economics*, 3(1), 10-15.
- Xue, X., Yao, H. D., & Davidson, L. (2022). Synthetic turbulence generator for lattice Boltzmann method at the interface between RANS and LES. *Physics of Fluids*, 34(5).
- Kim, E. H., Wang, Z., Zong, H., Jiang, Z., Fu, Z., & Pedrycz, W. (2023). Design of tobacco leaves classifier through fuzzy clustering-based neural networks with multiple histogram analyses of images. *IEEE Transactions on Industrial Informatics*, 20(3), 4698-4709.
- Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. *Mathematics*, 13(17), 2740.
- Li, W. W., Ziser, Y., Xie, Y., Cohen, S. B., & Ma, T. (2025, July). Tsprank: Bridging pairwise and listwise methods with a bilinear travelling salesman model. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1* (pp. 707-718).
- Li, W. W., Kim, H., Cucuringu, M., & Ma, T. (2026, August). Can LLM-based financial investing strategies outperform the market in long run?. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1* (pp. 2711-2722).

- Ma, Y., Qu, D., & Wang, Y. (2026). Tracking evolving communities in fake news cascades using temporal graphs. *Scientific Reports*. (<https://www.nature.com/articles/s41598-026-35175-4>)
- Lo, S. C., Zingaro, A., McCullough, J. W., Xue, X., Gonzalez-Martin, P., Joo, B., ... & Coveney, P. V. (2025). A multi-component, multi-physics computational model for solving coupled cardiac electromechanics and vascular haemodynamics. *Computer Methods in Applied Mechanics and Engineering*, 446, 118185.
- Li, W. W., & Ma, T. (2025). Learn to Rank Risky Investors: A Case Study of Predicting Retail Traders' Behaviour and Profitability. *ACM Transactions on Information Systems*, 44(1), 1-33.
- Hu, Z., Hu, M., & Li, Z. (2026). DynSupplyNet: A Dynamic Graph Neural Network with Temporal Fusion for Supply Chain Risk Prediction and Propagation Analysis. *Journal of Computer Science and Frontier Technologies*, 3(2), 1-13.
- Hu, L., & Shen, Y. (2026). A predictive analytics approach for forecasting global stock index returns using deep learning techniques. *Decision Analytics Journal*, 100685.
- Xue, Xiao, et al. "Equivariant U-Shaped Neural Operators for the Cahn-Hilliard Phase-Field Model." *arXiv preprint arXiv:2509.01293* (2025).
- Li, X., Li, Z., & Lin, X. (2025, September). Automated Implementation of Machine Learning-Based Causal Inference in Product Operations Decision Making. In *Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems* (pp. 204-209).
- Dong, J., Moayedi, R. Z., Ong, Y. S., & Moosavi-Dezfooli, S. M. (2026). Allies teach better than enemies: Inverse adversaries for robust knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.