Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Improved Adversarial Training Stability in Complex Visual Tasks

Emily Carter*[1], *Thomas Whitfield*[2], *Charlotte Bennett*[3]

Department of Computer Science, University College London, London WC1E 6BT, United Kingdom

***Corresponding author:** c.bennett.vision@ucl.ac.uk*

Abstract

Adversarial training is often used to improve the robustness of visual models. However, it may cause unstable gradients, large loss changes, and model collapse in complex visual tasks. This study presents a stable adversarial training method using dynamic loss weighting, spectral normalization, and gradient penalty. The dynamic loss weight controls the role of clean and adversarial samples during training. Spectral normalization limits model sensitivity, and gradient penalty reduces excessive parameter updates. Experiments were conducted on CIFAR-10 and an ImageNet subset under FGSM and PGD attacks. The perturbation strengths were set to 8/255 and 16/255. Compared with standard adversarial training, the proposed method improved robust accuracy by 6.5%–11.2%, reduced training loss fluctuation by about 34%, and improved FID by about 7.8%. These results show that the method can make adversarial training more stable while keeping useful visual features. This study provides a practical training strategy for robust visual learning in safety-related tasks, such as medical imaging, autonomous driving, and industrial inspection..

Keywords

adversarial training, visual learning, robust accuracy, gradient penalty, spectral normalization, dynamic loss weighting, training stability

INTRODUCTION

Deep neural networks have achieved strong performance in visual recognition tasks, but they remain sensitive to small adversarial perturbations. These perturbations are usually imperceptible to humans, yet they can mislead a model into producing incorrect predictions. This issue is particularly important in autonomous driving, medical imaging, industrial inspection, and other safety-sensitive scenarios, where prediction errors may lead to serious consequences [1,2]. In medical image reconstruction, differences in sampling strategy, reconstruction model design, and data conditions can also affect the stability and reliability of visual analysis systems, indicating that robustness is a key requirement for practical medical imaging applications [3]. Therefore, improving the robustness of deep neural networks against adversarial interference has become an important problem in reliable visual intelligence.

Adversarial training is one of the most widely used methods for improving model robustness. It introduces adversarial samples into the training process and encourages the model to learn more stable decision boundaries [4,5]. Compared with standard empirical risk minimization, adversarial training directly exposes the model to worst-case perturbations within a constrained input space [6,7]. This process can improve resistance to common attacks and reduce the model's dependence on non-robust visual features. Recent studies have further improved adversarial training through data augmentation, model averaging, synthetic data generation, robust distillation, and feature regularization [8,9]. These studies show that adversarial robustness is closely related to attack strength, training stability, feature smoothness, data diversity, and the regularity of the learned decision surface [10,11]. However, adversarial training still faces several practical limitations. A major problem is training instability. FGSM-based adversarial training is computationally efficient, but it may suffer from sudden robustness collapse during training [12,13]. PGD-based adversarial training is generally more reliable because it uses stronger iterative attacks, but it requires higher computational cost and may still experience robust overfitting after learning-rate decay [14,15]. Prior studies also indicate that loss fluctuation, sharp loss regions, unstable gradients, and large parameter updates can weaken adversarial training and reduce the final robust accuracy [16,17]. These problems suggest that improving robustness is not only a matter of generating stronger adversarial samples, but also requires better control of the optimization process. Model architecture also has a clear influence on adversarial robustness. Residual blocks, normalization layers, feature recalibration modules, vision transformers, and vision-language models can all affect the model's response to adversarial perturbations [18,19]. Some architectures improve feature representation and gradient flow, while others may increase sensitivity to small input changes. Although recent methods have reported promising results, many studies are still mainly evaluated on small-scale datasets such as CIFAR-10, CIFAR-100, and Tiny-ImageNet. Their effectiveness on larger ImageNet subsets and more complex visual scenes remains insufficiently verified. In addition, many papers focus mainly on robust accuracy, but provide limited analysis of loss fluctuation, gradient stability, and model collapse during training. Some methods also rely on extra data, heavy augmentation, or costly adversarial example generation, which limits their practical use in resource-constrained settings [20,21]. To address these gaps, this study proposes a stabilization method for adversarial training in complex visual recognition tasks. The proposed method combines dynamic loss weighting, spectral normalization, and gradient penalty within a unified training framework. The adaptive loss weight adjusts the relative contribution of clean and adversarial samples during training, so that the model can maintain standard accuracy while improving robustness. Spectral normalization is used to control the sensitivity of network layers and reduce unstable feature amplification. Gradient penalty further constrains excessive parameter updates and helps smooth the optimization process. Experiments are conducted on CIFAR-10 and an ImageNet subset under FGSM and PGD attacks, with perturbation strengths of 8/255 and 16/255. Robust accuracy, loss fluctuation, and FID are used to evaluate both adversarial resistance and training stability. The purpose of this study is to reduce robustness collapse, improve optimization stability, and enhance the practical reliability of adversarial training. Its significance lies in providing a lightweight and stable robustness-improvement strategy that does not depend heavily on extra data or overly expensive attack generation, making it more suitable for safety-sensitive visual applications where model stability and prediction reliability are both required.

2.Materials and Methods

2.1 Samples and Study Objects

Two image datasets were used in this study. CIFAR-10 was used as the basic dataset. It contains 60,000 color images from 10 classes, including 50,000 training images and 10,000 test images. An ImageNet subset was used to test the method in a more complex setting. This subset included 100 object classes, 120,000 training images, and 5,000 validation images. CIFAR-10 images were kept at 32 × 32 pixels. ImageNet-subset images were resized to 224 × 224 pixels. The study object was a convolutional visual model trained with adversarial samples. The experiment mainly examined whether dynamic loss weighting, spectral normalization, and gradient penalty could improve training stability under FGSM and PGD attacks.

2.2 Experimental Design and Control Settings

Four groups were designed for comparison. The baseline group used standard adversarial training without extra stability control. The second group used adversarial training with spectral normalization. The third group used adversarial training with gradient penalty. The proposed group used dynamic loss weighting, spectral normalization, and gradient penalty together. This setting was used to test the separate and combined effects of each module. FGSM and PGD were used to generate adversarial samples. The perturbation strengths were set to 8/255 and 16/255. All groups used the same network structure, optimizer, batch size, training epochs, and learning-rate schedule. Clean images and adversarial images were both used during training. This setting allowed a fair comparison between clean accuracy and robust accuracy

2.3 Measurement Methods and Quality Control

Model performance was measured by clean accuracy, robust accuracy, training loss fluctuation, gradient norm, and FID. Clean accuracy was calculated with original test images. Robust accuracy was calculated with adversarial test images generated by FGSM and PGD. Training stability was measured by recording loss values and gradient norms at each epoch. FID was used to measure changes in visual feature quality. For quality control, the training, validation, and test sets were fixed before the experiment. Test images were not used during training. Each experiment was repeated three times with different random seeds. The mean and standard deviation were reported. Attack settings, input normalization, data augmentation, and model initialization were kept the same across all groups.

2.4 Data Processing and Model Formulas

All images were normalized by channel mean and standard deviation. Random cropping and horizontal flipping were used only during training. Each training batch contained clean images and their adversarial versions. The total loss included clean loss, adversarial loss, and gradient penalty:

$$L_{total}=\alpha_t L_{clean}+(1-\alpha_t)L_{adv}+\lambda L_{gp}$$

where L_{clean} is the loss from clean images, L_{adv} is the loss from adversarial images, L_{gp} is the gradient penalty, λ is the penalty weight, and α_t is the dynamic weight at epoch ttt. Robust accuracy was calculated as:

$$RA=\frac{N_{correct}^{adv}}{N_{total}}\times 100\%$$

where $N_{correct}^{adv}$ is the number of correctly classified adversarial images, and N_{total} is the total number of test images. Loss fluctuation was calculated as the standard deviation of training loss across epochs.

2.5 Statistical Analysis and Model Evaluation

All results were reported as mean ± standard deviation from three runs. The proposed method was compared with the baseline and ablation groups. The comparison included clean accuracy, robust accuracy, loss fluctuation, gradient norm, and FID. The final model was selected based on

validation robust accuracy instead of training accuracy. Training curves were also checked to identify loss spikes, gradient explosion, and robustness collapse. A model was considered more stable when it showed higher robust accuracy, lower loss fluctuation, smaller gradient-norm variation, and no clear collapse in the late training stage.

3.Results and Discussion

3.1 Robust Accuracy Under FGSM and PGD Attacks

The proposed method improved robust accuracy under both FGSM and PGD attacks. Compared with standard adversarial training, robust accuracy increased by 6.5%–11.2% on CIFAR-10 and the ImageNet subset. The improvement was larger when the perturbation strength was 16/255. This result shows that the method worked better under stronger attacks. The baseline model kept good clean accuracy, but its robust accuracy dropped clearly under PGD. This drop was stronger on the ImageNet subset. This finding is consistent with previous studies. Weakly controlled adversarial training can perform well under simple attacks but may fail under stronger multi-step attacks [22,23]. The proposed method reduced the gap between FGSM and PGD results. This means that the model did not only learn patterns from one attack type. Gao, G. et al. (2026) also reported that FGSM-based training may show sudden robustness collapse when the decision boundary becomes unstable. Our results show that dynamic loss weighting and gradient control can reduce this problem in complex visual tasks. Fig. 1 shows the robust accuracy comparison.

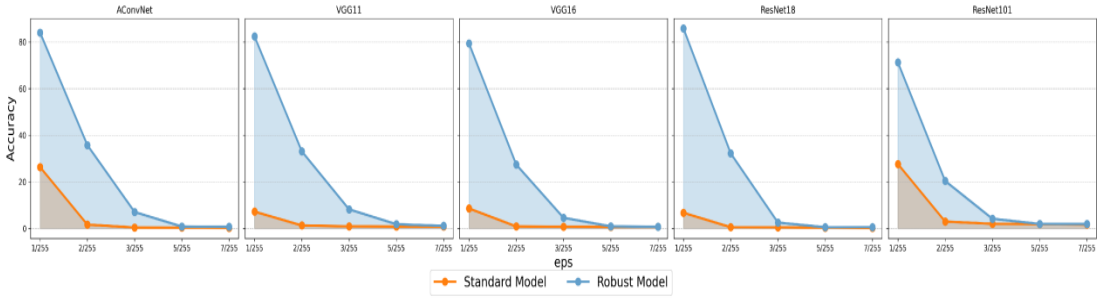
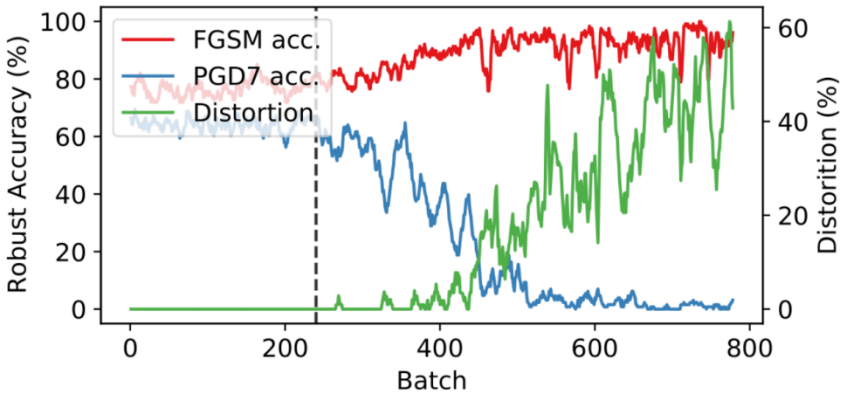


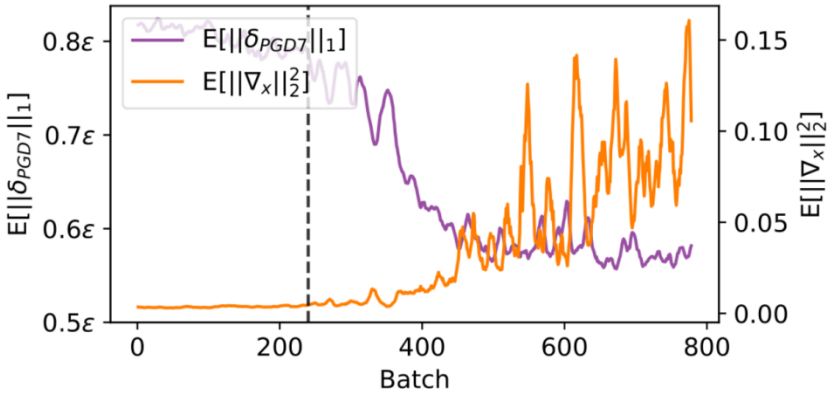
Fig. 1. Robust accuracy of each training method under FGSM and PGD attacks on CIFAR-10 and the ImageNet subset.

3.2 Training Stability and Loss Change

The proposed method made the training process more stable. Loss fluctuation decreased by about 34% compared with the baseline. This result shows that the model update became smoother after using dynamic loss weighting, spectral normalization, and gradient penalty [24,25]. In the baseline group, the loss curve showed clear spikes after adversarial samples were added. These spikes were related to large gradient changes and unstable responses to perturbed images. Spectral normalization reduced part of the fluctuation, but the effect was limited when it was used alone. Gradient penalty gave better control of sharp updates, but it could not fully balance clean loss and adversarial loss. The complete method showed the most stable curve because it controlled both loss weight and update size. This finding agrees with Wang, Y. et al. (2026), who linked sudden gradient changes with catastrophic overfitting in fast adversarial training. Fig. 2 shows an example of this instability.



(a) Robust accuracy and distortion



(b) Mean of absolute value of PGD7 perturbations and L_2 norm of the gradients of the images

Fig. 2. Training loss changes of each adversarial training method during different epochs.

3.3 Role of Each Stability Module

The ablation results showed that each module had a different role. Spectral normalization mainly reduced model sensitivity by limiting layer weight scale. It improved PGD robustness, but it had a smaller effect on loss fluctuation. Gradient penalty had a stronger effect on training smoothness because it limited sharp changes in the loss surface. However, when gradient penalty was used alone, the model still had difficulty keeping a good balance between clean accuracy and robust accuracy. Dynamic loss weighting helped solve this problem. It reduced the effect of adversarial loss in the early stage and increased it after the model became stable. The complete method achieved the best overall result. Compared with methods that depend on extra data, heavy augmentation, or longer adversarial training, this method improved robustness by making the training process more stable [26,27]. This design is more practical for complex visual tasks with limited data or computing resources.

3.4 Visual Feature Quality and Generalization

The FID improved by about 7.8%, which shows that the proposed method did not weaken the learned visual features. This point is important because adversarial training often improves robustness but reduces clean accuracy or feature quality. On CIFAR-10, the proposed method kept

clean accuracy close to the baseline and improved robust accuracy. On the ImageNet subset, clean accuracy decreased slightly, but robust accuracy improved more clearly. This trade-off is acceptable because the ImageNet subset has more class variation and background noise. Many previous studies mainly report final classification accuracy. In contrast, this study also examined loss fluctuation, gradient change, and FID. These indicators give a clearer view of training stability. Overall, the results show that adversarial robustness should not be judged only by final accuracy [28]. A stable method should also reduce loss oscillation, control gradient changes, and keep useful visual features during training.

4.Conclusion

This study developed a stable adversarial training method for complex visual tasks. The method uses dynamic loss weighting, spectral normalization, and gradient penalty to control model updates during training. The results showed that the method improved robust accuracy under FGSM and PGD attacks, reduced loss fluctuation, and decreased the risk of model collapse. Compared with standard adversarial training, the proposed method gave a smoother training process and kept useful visual features. Its main contribution is that it improves robustness by stabilizing training, rather than only increasing attack strength or using more data. This method has potential value in medical imaging, autonomous driving, industrial inspection, and intelligent monitoring, where stable prediction is important. However, this study still has limitations. The experiments were mainly based on CIFAR-10 and an ImageNet subset. More tests are needed on larger datasets and other visual tasks, such as object detection and semantic segmentation. Future work should also test the method with different model structures, attack types, and real-world visual scenes.

References

Chen, F., Li, S., Liang, H., Xu, P., & Yue, L. (2025). Optimization Study of Thermal Management of Domestic SiC Power Semiconductor Based on Improved Genetic Algorithm.

Kothuri, S. R., & Thaoroijam, K. (2026). A Robust Deep Ensemble Architecture for Safety-Critical Decision-Making in Autonomous Systems. Transactions on Emerging Telecommunications Technologies, 37(2), e70333.

Yan, B. (2026). A Comparative Study: Greedy versus GAN-Based Compressed Sensing for Fast MRI Reconstruction. International Journal of Advance in Applied Science Research, 5(3), 36-42.

Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artificial Intelligence Review, 58(1), 12.

Liang, R., Feifan, F. N. U., Liang, Y., & Ye, Z. (2025). Emotion-Aware Interface Adaptation in Mobile Applications Based on Color Psychology and Multimodal User State Recognition. Frontiers in Artificial Intelligence Research, 2(1), 51-57.

Mohan, A., & Schön, T. (2026). Towards Robust Agents: A Survey of Adversarial Attacks and Defenses in Deep Reinforcement Learning. IEEE Access.

Zhu, W., Yao, Y., & Yang, J. (2025). Optimizing Financial Risk Control for Multinational Projects: A Joint Framework Based on CVaR-Robust Optimization and Panel Quantile Regression.

Gordienko, Y., Nowakowski, G., Kochura, Y., Taran, V., & Stirenko, S. (2025, October). Generative Data Augmentation by Dataset Distillation. In International Workshop on Artificial Intelligence for Biomedical Data (pp. 105-118). Cham: Springer Nature Switzerland.

Feng, H. (2024, October). Design of Intelligent Charging System for Portable Electronic Devices Based on Internet of Things (IoT). In 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE) (pp. 568-571). IEEE.

Holla, H., Polepalli, S. R., & Sasikumar, A. A. (2025). Adversarial threats to cloud IDS: robust defense with adversarial training and feature selection. IEEE Access.

Liu, S., Liu, X., & Feng, H. (2025, November). Research on AI-Driven Visual Design and Immersive Interactive Experiences Based on Multimodal Cognition and User. In Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 734-740).

Madabhushi, S., & Dewri, R. (2025). Mitigating Over-Generalization in Anomalous Power Consumption Detection using Adversarial Training. ACM Transactions on Cyber-Physical Systems, 9(3), 1-25.

Chen, X., Xiao, H., Zeng, Z., Zhang, S., & Xiao, R. (2025). Fine-Grained Multivariate Time Series Anomaly Detection via Causal Inference. Knowledge-Based Systems, 114765.

Kose, K., & Zhou, B. (2025). Adversarial Training for Aerial Disaster Recognition: A Curriculum-Based Defense Against PGD Attacks. Electronics, 14(16), 3210.

Gao, X., & Lan, Y. (2025, December). Data-Driven Identification Method for Low-Voltage Distribution Station-User Relationships Based on Voltage Sequence Correlation. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 508-512). IEEE.

Gaur, S., Singh, A., Kansal, V., & Prakash, S. (2025, December). MART+ AWP: A Hybrid Adversarial Training Framework to Reduce Robust Overfitting. In 2025 International Conference on Decision Aid Sciences and Applications (DASA) (pp. 1684-1689). IEEE.

Gao, X., Chen, J., & Huang, M. (2026, January). Research on Risk Dependency Structures and Resource Allocation Optimization in New Energy Technology Collaboration within Enterprise Distributed Innovation. In 2026 International Conference on Electrical Engineering, Intelligent Control and Artificial Intelligence (EEICAI) (pp. 43-47). IEEE.

Hassija, V., Palanisamy, B., Chatterjee, A., Mandal, A., Chakraborty, D., Pandey, A., ... & Kumar, D. (2025). Transformers for vision: A survey on innovative methods for computer vision. Ieee Access.

Xu, S., Ma, Q., Liu, H., & Yue, L. (2026). Continuous Reorganization and Performance Preservation of Agent Memory Structure Under Distributed Change Environments.

Singh, R. K. (2025). Generative Adversarial Networks for Data Augmentation in Low-Resource NLP Tasks.

Liu, S., & Yim, J. (2025). Research on Generative AI Creation Systems Based on Visual Language Modeling: Human-Machine Collaboration and Cognitive Feedback Mechanisms. Available at SSRN 6139770.

Shreeya, K. N., Subburaj, B., Saketh, K. S. G., Padmavathy, T. V., Alphonse, S., & Subramanian, G. (2026). A quantum resilient deepfake detection framework using enhanced resnext and post quantum cryptography defence. Scientific Reports.

Gao, G., Gao, R., Gao, R., & Zhou, H. (2026). Engineering Analysis and Quantitative Research on the Platform-Based Evolution of Enterprise Communication Systems.

Dwivedi, P., Islam, B., & Kajal, M. (2025). Smooth gradient loss: a loss function for gradient regularization in deep learning optimization. The Journal of Supercomputing, 81(15), 1-45.

Wang, Y., Yin, X., Chen, J., & Wang, Y. (2026). Evidence-Based Study on Low-Burden Digital Phenotyping for Precision Screening of Oral Anti-Obesity Drug Efficacy.

Gu, X. (2026, January). Analysis of the Synergistic Role of User Natural Language Feedback and Behavioral Data in Adjusting Recommendation Strategies for Mental Health Platforms. In Proceedings of the 2026 5th International Conference on Big Data, Information and Computer Network (pp. 198-203).

Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). The Impact of Visual Language Strategies in Public Art Creation on Community Spatial Perception and Public Behavior.

Haref, Q. M., Long, J., & Yang, Z. (2025). Enhanced Video-Based Fall Detection Using 3D Human Pose Estimation and Advanced Generative Adversarial Networks: Integrating WGAN Data Augmentation with a Hybrid 3D CNN–GRU Architecture. KSII Transactions on Internet & Information Systems, 19(12).