2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Enhancing the Robustness of Classification Against Adversarial Attacks through a Dual-Enhancement Strategy

Guo Niu[1][0000-0002-1552-7310] , Shuaiwei Jiao[2](✉)[0009-0000-3805-323], Nannan Zhu[3][0000-0003-4038-3053], Juxin Liao[2][0009-0007-4371-3739], Shengjun Deng[2][0009-0003-0089-2651], Tao Li[2][0009-0006-6802-5521], Xiongfei Yao[2][0009-0009-8464-8448], and Huanlin Mo[2][0009-0006-6802-5521]

[1] School of Electronic and Information Engineering, Foshan University, No. 18 Jiangwan 1st Road, Foshan, 528225, Guangdong, China
[2] School of Computer Science and Artificial Intelligence, Foshan University, No. 18 Jiangwan 1st Road, Foshan, 528225, Guangdong, China
[2] School of Systems Science and Engineering, Sun Yat-sen University, No. 135 Xingang West Road, Guangzhou, 510275, China
fwind190@gmail.com

Abstract. Deep neural networks have achieved remarkable success in target classification, but as accuracy improves, model robustness has become a growing concern. Existing methods, such as adversarial training, enhance robustness, yet adversarial examples can still lead to high-confidence, incorrect predictions. To address this issue, we propose a new defense mechanism—Dynamic MixCut. This method combines the advantages of multi-box CutMix and Mixup by enhancing the diversity and complexity in the sample generation process, enabling more effective defense against complex adversarial attacks, especially in dynamic perturbation environments. Through in-depth theoretical analysis, we reveal the fundamental reasons behind the robustness limitations of traditional Mixup under multi-step attacks, particularly the limitations of mixing adversarial perturbations between samples. Furthermore, the Dynamic MixCut method enhances the model's adaptability to diverse attack strategies by integrating more sophisticated perturbation designs in the generation of adversarial examples, thereby mitigating the trade-off between standard accuracy and adversarial robustness. Experimental results on the CIFAR-10 and SVHN datasets demonstrate that the Dynamic MixCut method improves adversarial accuracy by over 10% on average compared to the baseline while preserving standard accuracy. This research provides novel insights into robust training for object classification tasks and contributes to the advancement of adversarial training techniques.

Keywords: Object classification, Adversarial Attacks, Multi-step Attacks, Adversarial Robustness, Robust Training.

1 Introduction

In recent years, deep learning technologies have made significant advancements in various fields such as computer vision and natural language processing [1]. Particularly, deep learning models have demonstrated superior performance over traditional methods in tasks like image classification, object classification, and semantic segmentation [2]. However, despite their outstanding performance across many tasks, deep learning models have been shown to be highly vulnerable. Small, imperceptible perturbations can be added to the original data, which are sufficient to cause the model to confidently make completely erroneous predictions. Numerous studies have confirmed the vulnerability of neural networks to adversarial examples and test inputs. Even slight, carefully crafted changes to test inputs are enough to cause misclassification by the model.

Researchers have attempted to address this issue through adversarial training. While most methods improve the model's resistance to attacks [3,4], they also degrade the model's standard test accuracy [5]. These methods generate adversarial samples by calculating the gradient of the loss function with respect to input samples, adding small perturbations to the input data, and introducing these adversarial examples into the training process. This allows the model to learn more robust decision boundaries under adversarial conditions. However, despite some success in improving robustness, these methods still have shortcomings, especially when facing complex multi-step attacks, where their robustness often falls short of the desired level.

Although these methods have made significant progress in enhancing adversarial robustness, how to balance robustness with standard accuracy remains an urgent problem. Especially when facing various types of adversarial attacks, the effectiveness of different methods in improving robustness is inconsistent. Designing more effective adversarial training strategies to improve model robustness in different attack scenarios, while maintaining performance under normal conditions, continues to be a challenge.

In this study, we propose a novel defense mechanism: a robust defense framework based on dynamic mix-cut fusion for object classification. We theoretically analyze the limitations of traditional Mixup in multi-step adversarial attacks and introduce a multi-box dual CutMix approach for adversarial training with soft label data augmentation. Our method aims to improve robustness by generating more effective adversarial samples while minimizing the trade-off between standard accuracy and adversarial robustness. Our contributions are as follows:

- We propose Multi-box Dynamic MixCut, a novel augmentation method for spatially adaptive region mixing, improving robustness against adversarial perturbations.
- Our analysis reveals Mixup’s vulnerability to adversarial attacks due to indiscriminate mixing of background and salient regions, and we address this by introducing a training framework that enhances robust generalization through Lipschitz continuity analysis.
- We analyze our proposed method with the results of experiments on CIFAR-10 [6] and SVHN [7] dataset, demonstrating that Dynamic MixCut significantly enhances the robustness of state-of-the-art adversarial training methods.

2 Related Works

2.1 Adversarial training

Early adversarial training methods, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) [8-10], generate adversarial samples by calculating the gradient of the loss function with respect to input samples and adding small perturbations to the input data. These methods perform well against single-step attacks but show poor effectiveness when confronted with more complex multi-step attacks.

In addition to FGSM and PGD, many other powerful adversarial attack methods effectively challenge existing defense mechanisms. For example, the Carlini-Wagner (CW) attack [11] employs an optimization strategy to generate highly imperceptible adversarial samples, minimizing the perturbation size while ensuring the model misclassifies, demonstrating strong attacking capabilities in high-dimensional data and complex models. In contrast, the SPSA attack [12] is a black-box attack method that uses randomized gradient estimates to generate adversarial samples, suitable for scenarios where the model gradients are inaccessible. The Square attack [13] generates highly effective and imperceptible adversarial samples by randomly optimizing perturbation regions in the image space, making it particularly effective for attacking high-dimensional data. Another emerging attack strategy is Feature Scattering [14], which perturbs or scatters key features in the feature space, forcing the model to misclassify on decision boundaries without relying on pixel-level perturbations in the image.

Although existing defense methods have improved adversarial robustness to some extent, they typically sacrifice standard accuracy in exchange for adversarial robustness, leading to a trade-off between accuracy and robustness [15, 16]. Robust generalization refers to the model's ability not only to defend against adversarial attacks seen during training but also to effectively handle unseen adversarial samples. Improving robust generalization is an important research direction in adversarial training, with the aim of allowing models to defend against unknown threats [17], rather than being limited to specific types of attacks.

2.2 Data Augmentation in Adversarial Training

Data augmentation has become a prevalent and effective strategy in adversarial training, with the primary goal of enhancing the model's robustness [18]. By incorporating a wider variety of training samples, data augmentation not only improves the model's generalization capabilities but also strengthens its resilience to adversarial attacks. Traditional image augmentation techniques, such as rotation, scaling, and flipping, have shown some benefits in boosting robustness; however, their impact on defending against adversarial perturbations remains limited. In recent years, augmentation methods based on mixing inputs, such as Mixup and CutMix, have emerged as powerful tools for improving a model's adversarial resilience.

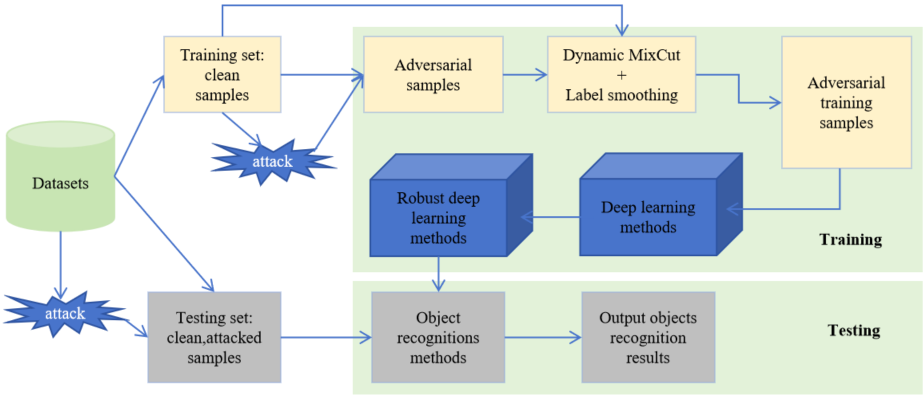


Fig. 1. Overview of the dual-enhancement strategy for improving classification robustness against adversarial attacks.

Several approaches have already been proposed, including ManifoldMixup [19], Un-Mix [20], Puzzle-Mix [21], SaliencyMix [22], TokenMix [23], and TrasMix [24]. These methods enhance robustness by introducing more sophisticated mixing strategies. Saliency-based methods, such as Attentive-CutMix and SaliencyMix, focus on identifying key image regions (e.g., objects or edges), thereby reducing interference from less important background areas and improving adversarial defense.

Building on these advancements, we propose DynaMixCut, a novel approach that combines the strengths of Mixup and CutMix. DynaMixCut effectively handles both global and local image variations, significantly enhancing robustness against complex adversarial attacks. By refining the generation of adversarial examples, DynaMixCut increases training data diversity and better captures the varying contributions of different image regions to the label space. This approach enables the model to learn more robust decision boundaries while minimizing the trade-off between standard accuracy and adversarial robustness.

3 Method

In this section, we theoretically explain the limitations of the traditional Mixup method when addressing multi-step adversarial attacks, particularly its shortcomings in defending against dynamic perturbations and complex attack strategies. By introducing the Dynamic MixCut method, which combines the strengths of CutMix and Mixup along with a multi-box mechanism, we effectively enhance the diversity and complexity of adversarial samples. This improvement significantly increases the model's adaptability to multi-step adversarial attacks. The method not only enhances the model's robustness in specific attack scenarios but also boosts its generalization ability, enabling the model to better withstand previously unseen attack strategies. As a result, the overall performance and robustness of the model are significantly improved. The main process is shown in Figure 1.

3.1 Theoretical Motivation

To enhance model robustness against adversarial attacks, the Mixup method increases the continuity between samples, enabling the model to better adapt to small perturbations. For any two samples (xi, yi) and (xj, yj), Mixup generates a new sample as:

$$\bar{\mathrm{x}} = \lambda \mathrm{x}_\mathrm{i} + (1-\lambda)\mathrm{x}_\mathrm{j}, \quad \bar{\mathrm{y}} = \lambda \mathrm{y}_\mathrm{i} + (1-\lambda)\mathrm{y}_\mathrm{j} \tag{1}$$

Where λ is sampled from a uniform distribution (U(0, 1). For single-step attacks such as the Fast Gradient Sign Method (FGSM), the adversarial loss is defined as:

$$L_{adv}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max_{||\delta_i||} l(\theta, (x_i + \delta_i, y_i)) \tag{2}$$

ϵ is the upper limit of disturbance. After applying Mixup, the adversarial loss becomes:

$$L_{adv}(\theta) = \sum_{i=1}^{n} l(\theta, (\bar{\mathrm{x}}, y_i)) \tag{3}$$

$\bar{\mathrm{x}}$ is a linear interpolation generated by the Mixup technique between the original image and its corresponding adversarial image. While Mixup demonstrates effectiveness in mitigating adversarial loss under single-step attacks (e.g., Fast Gradient Sign Method, FGSM), its robustness significantly deteriorates when confronted with more sophisticated multi-step attacks, such as Projected Gradient Descent (PGD). The adversarial loss for PGD is defined as:

$$L_{adv}(\theta) = \sum_{i=1}^{n} l(\theta, (x_i + \delta_i, y_i)) \tag{4}$$

Due to the dynamic nature of PGD perturbations in magnitude and direction, Mixup-generated samples often fail to defend against such attacks, leading to reduced model robustness. To further analyze Mixup's limitations, we perform a second-order Taylor expansion of the adversarial loss [25]:

$$L_{adv}(\theta) \approx l\big(\theta, (x, y)\big) + \nabla l(\theta, (x, y)^T \delta + \frac{1}{2}\delta^T H \delta \tag{5}$$

Where H is the Hessian matrix of the loss function with respect to the input, capturing second-order derivative information, and δ is the adversarial perturbation.

While Mixup reduces adversarial loss under small perturbations, it struggles to capture second-order changes in the loss function under larger perturbations, particularly in multi-step attacks. As sample complexity increases, Mixup's adversarial perturbation effect diminishes. Assuming a constant c_x such that $||x_i|| \geq c_x\sqrt{d}$ for all i, where x_i is the input sample, d is the feature dimension, and $||x_i||$ is theL_2 norm, the perturbation magnitude generated by Mixup is proportional to the sample's complexity and feature dimension d. This leads to the inequality:

$$L_{mix}(\theta) \geq \frac{1}{n}\sum_{i=1}^{1} l_{adv}(\epsilon_{mix}\sqrt{d}, (x_i, y_i) \tag{6}$$

This suggests that Mixup's effectiveness weakens as sample complexity increases, the adversarial perturbation effect of the Mixup method weakens, indicating the need to explore potential methods combined with adversarial training.

3.2 Dynamic MixCut

To enhance the model's robustness against adversarial attacks, we propose the Dynamic MixCut algorithm. This algorithm uniquely combines dynamic mixing techniques to generate a rich variety of sample combinations, effectively improving the model's resistance to complex attacks. Additionally, Dynamic MixCut enhances the diversity of samples in adversarial training, making the model's performance more stable and reliable across different scenarios.

CutMix operation for generating adversarial samples. After generating the mixed samples, we introduce the Mixup strategy to further improve sample diversity. The Mixup technique generates new samples by linearly interpolating the features of two samples, thereby increasing the continuity of the samples and the model's adaptability to small perturbations. Specifically, for the mixed samples $\tilde{x}_1$ and $\tilde{x}_2$, we use weight coefficients β_1 and β_2 to generate the final input sample $\bar{x}$:

$$\bar{\mathrm{x}} = \beta_1 \cdot \tilde{x}_1 + \beta_2 \cdot \tilde{x}_2 \tag{7}$$

Here, β_1 and β_2 are sampled from the uniform distribution U(0, 1). With the Mixup operation, the model receives samples durings training that are interpolated between the original and adversarial samples, increasing the diversity of the training data.

Additionally, we apply smoothing techniques to the labels to prevent the model from overfitting to them.

Targeted Attacks and Mixup for Robust Generalization. Neural networks exhibit local linearity, which forms the theoretical foundation for improving adversarial robustness. Data augmentation techniques like Mixup exploit this property, as illustrated in Figure. 2, where the model learns to interpolate linearly between neighboring samples by reinforcing strong linearity in local input regions, thereby expanding the input sample space and enhancing resistance to perturbations. The local linearity hypothesis for adversarial noise suggests that a model’s output can be approximated as a linear transformation of the input during training, providing intrinsic resistance to adversarial attacks.

According to Lipschitz continuity, reducing the Lipschitz constant decreases the model’s sensitivity to input perturbations, thereby improving adversarial robustness. Under this condition, the model’s response to perturbations remains smooth within a local region, ensuring that Mixup-generated samples through interpolation experience limited perturbations, thus enhancing robustness against adversarial noise.

```
Algorithm 1 Dynamic MixCut Adversarial Training
Require: Training set D, model parameters $\theta$, adversarial sample generation function G, loss function $l$, number of iterations K, smoothing factor $\alpha$. and β₁ generated from uniform(0, 1) distribution.
for each epoch =1 to N do
    for each training sample lable pair (x,y) ~ D do
        Generate two distinct adversarial samples:
            $p_1 \leftarrow G(x, y; \theta)$
            $p_2 \leftarrow G(x, y; \theta)$
        for k=1 to K do
            Generate random box coordinates for CutMix: $a_k$
            Perform CutMix operation:
                $\tilde{x}_1 \leftarrow (1 - a_k)x + a_k p_1$
                $\tilde{x}_2 \leftarrow (1 - a_k)x + a_k p_2$
        end for
        Mixup among blended samples:
            $\bar{x} = \beta_1 \cdot \tilde{x}_1 + (1 - \beta_1) \cdot \tilde{x}_2$
        Label Smoothing:
            $\bar{y} \leftarrow y \cdot (1 - \alpha) + \frac{\propto}{classes}$
        Compute the loss: $l(\bar{x}, \bar{y}; \theta)$
        Backpropagate the loss and update parameters $\theta$
    end for
end for
Output: Robust model parameters $\theta$
```

Targeted and untargeted attacks behave differently within this framework. Targeted attacks can manipulate the model's output within a local input region, driving predictions toward a specific class. For example, an attacker can modify a sample x with perturbation ϵ to make $F(x + \epsilon) = y_i$ the dominant output, thereby achieving the attack. Compared to untargeted attacks, targeted attacks generate more diverse adversarial samples, improving the model's overall robustness. Therefore, adversarial training with random targeted attacks enhances the robustness of Mixup-trained models.

Training Process of the Combined Dynamic MixCut. The entire training process of Dynamic MixCut is iterative. In each epoch, we first generate and mix adversarial samples for each sample (x, y) in the training set D. The number of adversarial samples generated and the number of Mixup operations can be adjusted according to experimental settings to meet the needs of different datasets and tasks.

In each iteration, two independent adversarial samples p_1 and p_2 are first generated, followed by multiple CutMix operations to obtain $\tilde{x}_1$ and $\tilde{x}_2$. Then, these mixed samples are combined using Mixup to get the final training sample $\bar{x}$, and its corresponding label $\bar{y}$ is computed. The model is then trained based on these mixed samples, and the model parameters θ are updated by minimizing the cross-entropy loss.

The key to this process is that by generating and mixing adversarial samples multiple times, we provide the model with more challenging training samples, which improves its robustness when facing adversarial attacks. Moreover, the dynamic nature of Dynamic MixCut allows the model to adaptively adjust the sample generation and mixing strategies at different stages of training, further enhancing its performance.

Therefore, by combining the advantages of CutMix and Mixup, the corresponding algorithm is as follows in Algorithm 1.

4 Experimental Results and Analysis

4.1 Experiment Settings

Dataset. To verify the effectiveness of our method, we use two widely used public datasets: SVHN (Street View House Numbers) and CIFAR-10. The SVHN dataset contains 732,577 training images and 26,032 test images, all of which are 32x32 pixels in size and cover 10 digit categories. Its complex backgrounds and diverse digit styles make it an ideal benchmark for testing model robustness. The CIFAR-10 dataset consists of 60,000 32x32 pixel color images, divided into 10 categories (e.g., airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck), with 6,000 images per category. Known for its simple backgrounds and standardized images, CIFAR-10 is a classic benchmark for evaluating model performance.

By conducting experiments on these datasets, we comprehensively assess the performance and robustness of our model. Note that PGD and CW attacks with iteration steps are denoted as PGDT and CWT, respectively, while the original test set is referred to as Clean.

Implementation Details. We employ the ResNet18 [26] model for robust training on the CIFAR-10 and SVHN. During training, we use SGD as the optimizer with a momentum of 0.9 and a weight decay of 2e-4. All loss functions are based on cross-entropy loss. The optimizer relies heavily on learning rate adjustments, and we adopt a warm-up strategy followed by exponential decay to gradually reduce the learning rate.

For CIFAR-10, we set the initial learning rate to 0.1, warming up to 1 over 10 epochs and then multiplying by 0.985 after each epoch. For SVHN, we use the same settings: an initial learning rate of 0.032, warming up to 1 over 10 epochs and decaying by 0.985 after each epoch. We set the total number of epochs to 400 and the batch size to 128 for all datasets. The adversarial examples used in training are generated by PGD-10, with an l_∞ norm $\frac{8}{255}$. To validate the effectiveness of our method, we use FGSM, PGD, and CW for adversarial testing. We used a computer with Ubuntu20.04, GeForce RTX4090, python-3.9.19 and Torch-2.1.1+cu118, and mainly compared the following settings in our experiments:

- Standard: Models trained with the original dataset.

- PGD: Models trained with adversarial examples from PGD, step size = 2, iteration steps = 10.
- TRADES: Designed to balance model accuracy with adversarial robustness by introducing a regularization term to enhance adversarial sample training.
- Feature Scatter: Optimizes model performance across different data distributions by analyzing feature distribution scatter, helping to identify potential overfitting and underfitting issues.
- Our Method: We apply our proposed method to models based on PGD.

4.2 Ablation Studies

Effect of CutMix and Targeted Attacks We first analyze the impact of using original samples versus our proposed method on model performance during training. To validate the effectiveness of mixed attacks (CutMix and targeted attacks), we conduct ablation experiments comparing standard adversarial training with our approach. As shown in Table. 1, both CutMix and targeted attacks simultaneously enhance the model's robustness and accuracy under cross-entropy loss.

Table 1. Ablation study on target attacks and CutMix effects, showing robustness metrics under clean/adversarial settings (PGD-20/100 steps) with Lipschitz constants.

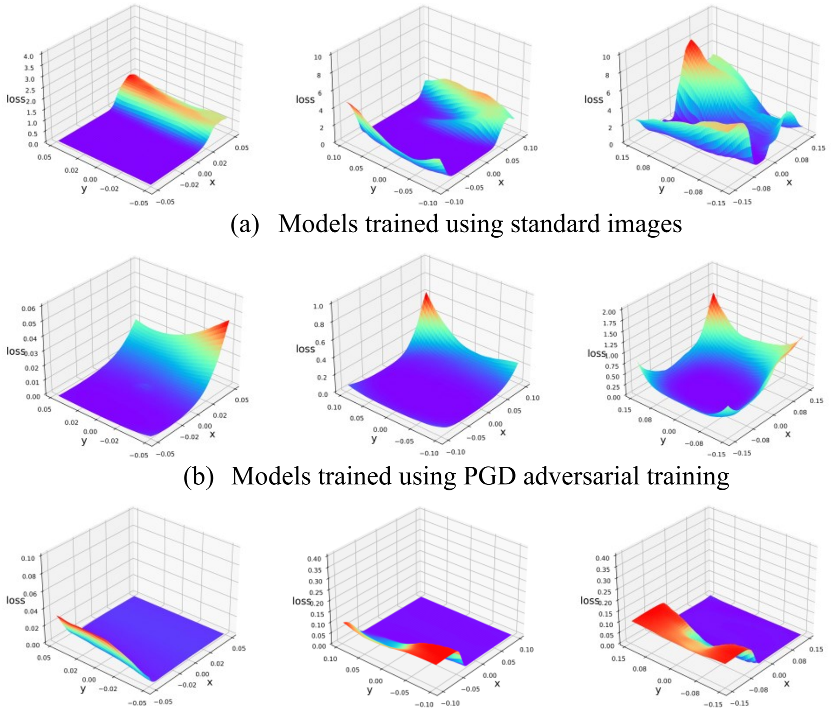
Method\Attack	**Targeted**	Clean	PGD20	PGD100	Lipz
Without CutMix	-	81.5	44.3	43.9	71.4
	√	86.4	40.1	39.6	133.9
Dynamic MixCut	-	86.3	59.1	56.7	1.63
	√	89.8	67.7	64.5	1.35

Table 2. Impact of box count on model performance: natural accuracy and PGD-20 results on SVHN dataset.

Dataset	Box Number	Natural Accuracy			PGD-20		
		Best	Final↓	Diff	Best	Final	Diff↓
SVHN	One Box	96.1	95.3	0.8	84.2	69.1	15.1
	Two Box	96.5	96.0	0.5	86.9	79.3	7.6
	Three Box	96.5	95.9	0.6	85.3	71.1	14.2
	Four Box	96.6	96.0	0.6	86.9	73.4	13.5
	Five Box	96.7	96.0	0.7	86.4	74.3	12.1
	Six Box	96.7	96.1	0.6	88.7	77.9	10.8

Effect of Box Count. Finally, we study the impact of the number of boxes on the results. By adjusting the number of boxes used in each input image, we analyzed the effect of different box counts during the training process. Our experiments showed that the number of boxes directly affects the model's performance, particularly when facing adversarial attacks. Increasing the number of boxes appropriately can provide more contextual information and details, which helps improve the model's classification ability in complex scenarios. However, using too many boxes may lead to information overload, thereby affecting the model's learning efficiency. Therefore, exploring the effect of the number of boxes on experimental results is crucial. The specific experimental results are shown in Table. 2.

Loss Landscape. To investigate the impact of adversarial examples on model outputs, we visualize the loss landscapes of different models. These loss surfaces capture the variation in model loss within the neighborhood of a single sample. We define two orthogonal directions: the attack direction relative to the sample as the x-axis and a random direction orthogonal to the x-axis as the y-axis, forming a plane in the sample space. We then sample points on this plane in a grid, calculating the loss at each grid, calculating the loss at each grid point. The difference between the grid loss and the original loss is used as the z-axis value, and the resulting surface is plotted for visualization.



(a) Models trained using standard images

(b) Models trained using PGD adversarial training

(c) Models trained using Our Method adversarial training

Fig. 3. Comparision of adversarial training methods under different attack radius:0.05,0.1,0.

The loss surfaces are plotted at different scales: the smallest scale (0.05 attack radius) captures fine-grained variations, while the largest scale (0.15 attack radius) reflects broader trends. A commonly used attack radius is $\frac{8}{255}$, which falls within the smallest scale of our plots. If the loss variation within this range is minimal, it indicates that the model exhibits excellent smoothness and robustness to attacks. Larger scales further demonstrate the model's smoothness under stronger perturbations. However, when the attack radius exceeds 0.15, adversarial samples severely degrade the semantic information of the image, rendering further analysis meaningless.

4.3 Effectiveness of the Proposed Algorithms

In this section, we demonstrate the robust generalizability of our adversarial training method. The experiments are carried out on the SVHN and CIFAR-10 datasets, using the ResNet18 model as the base architecture. The training settings follow the configurations described in Section 4.1, including the use of SGD with a warm-up phase and exponential decay for the adjustment of the learning rate.

To evaluate the effectiveness of our method, we test against a variety of adversarial attack algorithms, including FGSM, PGD-20, SPSA, CW-20. We report the accuracy on clean test images and under adversarial attacks, as well as the model's adversarial Lipschitz continuity. The perturbation bound is set to $\epsilon = \frac{8}{255}$ with a step size of $\frac{2}{255}$, consistent with standard settings in the literature.

Dataset	Method	Nat	FGSM	PGD20	SPSA	CW20	Lipz
CIFAR-10	Standard	96.4	20.6	0.0	0.0	0.0	566.7
	PGD_untarget	81.4	52.4	44.6	54.0	44.5	7.1
	TRADES	82.0	57.9	52.2	58.7	50.0	23.5
	Feature Scatter	91.4	73.7	57.5	52.5	49.8	8.4
	Our Method	89.5	74.8	70.4	76.5	82.8	1.27
SVHN	Standard	96.4	30.3	0.2	0.9	0.2	438.1
	PGD_untarget	91.4	66.2	52.5	58.9	49.8	52.1
	TRADES	91.2	69.7	58.8	63.1	54.9	19.4
	Feature Scatter	93.4	72.1	62.6	69.1	60.1	15.1
	Our Method	95.2	82.0	79.6	83.0	90.7	1.29

Table. 3. Comparison of Model Performance with and without Dynamic MixCut under Different Attacks (Clean, FGSM, PGD-20, SPSA and CW-20), and the Lipschitz constant of the model was calculated.

The results, summarized in Table. 3, show that our fusion algorithm outperforms the baseline methods in almost all evaluation metrics. Specifically, our method achieves higher robustness against adversarial attacks while maintaining competitive accuracy on clean data. These results highlight the effectiveness of our fusion algorithm in enhancing adversarial robustness across diverse datasets and attack scenarios. By combining the strengths of CutMix and Mixup, our approach provides a more comprehensive defense mechanism against complex adversarial perturbations.

5 Conclusion

In this paper, we propose Dynamic MixCut for improving the robustness of object classification. This approach integrates the strengths of CutMix and Mixup with a multi-box mechanism to enhance perturbation diversity during training. Dynamic MixCut addresses the limitations of traditional Mixup in defending against multi-step adversarial attacks, particularly its vulnerability to dynamic perturbations. Through theoretical analysis, we identify that Mixup's tendency to overfit to adversarial features and its limited adaptability to perturbation variations undermine its robustness. Dynamic MixCut mitigates these issues by refining the adversarial sample generation process, significantly reducing the trade-off between standard accuracy and adversarial robustness.

Experimental results demonstrate that Dynamic MixCut outperforms existing defense methods, achieving superior robustness against complex adversarial attacks on CIFAR-10, and SVHN datasets. While our method advances adversarial robustness, future work should focus on improving traditional data augmentation techniques for multi-step attacks and dynamic perturbation environments. By integrating more flexible perturbation generation strategies and adaptive training mechanisms, data augmentation can further strengthen adversarial training.

References

1. Kasri, W., Himeur, Y., Alkhazaleh, H.A., Tarapiah, S., Atalla, S., Mansoor, W.,Al-Ahmad, H.: From vulnerability to defense: The role of large language models in enhancing cybersecurity. Computation (2), 2079–3197 (2025)
2. Liu, C., Dong, Y., Xiang, W., Yang, X., Su, H., Zhu, J., Chen, Y., He, Y., Xue,H., Zheng, S.: A comprehensive study on robustness of image classification models:Benchmarking and rethinking. International Journal of Computer Vision (2), 567–589 (2025)
3. Zühlke, M.M., Kudenko, D.: Adversarial robustness of neural networks from theperspective of lipschitz calculus: A survey. ACM Computing Surveys (6), 1–41(2025)
4. Xie, T., Dai, K., Wang, K., Li, R., Zhao, L.: Deepmatcher: a deep transformer-based network for robust and accurate local feature matching. Expert Systems with Applications p. 121361 (2024)
5. Goodfellow, arkehabadi, A., Oftadeh, P., Shafaie, D., Hassanpour, J.: On the connection between saliency guided training and robustness in image classification. In: 2024 12th International Conference on Intelligent Control and Information Processing (ICICIP). pp. 203–210 (2024)
6. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images.Handbook of Systemic Autoimmune Diseases (2009)
7. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., et al.: Readingdigits in natural images with unsupervised feature learning. In: NIPS workshop on deep learning and unsupervised feature learning. p. 4 (2011)
8. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015)
9. Mądry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. stat (9) (2017)

10. Wang, Y., Cheng, S., Du, X.: Scs-jpgd: Single-channel-signal joint projected gradient descent. Applied Sciences pp. 2076–3417 (2025)
11. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In:2017 ieee symposium on security and privacy (sp). pp. 39–57 (2017)
12. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. pp. 506–519 (2017)
13. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European conferenceon computer vision. pp. 484–501 (2020)
14. Zhang, H., Wang, J.: Defense against adversarial attacks using feature scattering-based adversarial training. In: Advances in Neural Information Processing Systems(2019)
15. Liu, C., Huang, Z., Salzmann, M., Zhang, T., Süsstrunk, S.: On the impact of hard adversarial instances on overfitting in adversarial training. Journal of Machine Learning Research pp. 1–46 (2024)
16. Pan, C., Li, Q., Yao, X.: Adversarial initialization with universal adversarial perturbation: A new approach to fast adversarial training. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 21501–21509 (2024)
17. Aliferis, C., Simon, G.: Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and ai. Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls pp. 477–524 (2024)
18. Ribas, L.C., Casaca, W., Fares, R.T.: Conditional generative adversarial networks and deep learning data augmentation: A multi-perspective data-driven survey across multiple application fields and classification architectures. AI pp. 2673–2688(2025)
19. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In:International conference on machine learning. pp. 6438–6447 (2019)
20. Shen, Z., Liu, Z., Liu, Z., Savvides, M., Darrell, T., Xing, E.: Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2216–2224 (2022)
21. Kim, J.H., Choo, W., Song, H.O.: Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In: International conference on machine learning. pp. 5275–5285 (2020)
22. Walawalkar, D., Shen, Z., Liu, Z., Savvides, M.: Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In:ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3642–3646 (2020)
23. Choi, H.K., Choi, J., Kim, H.J.: Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. Advances in Neural Information Processing Systems pp. 14224–14235 (2022)
24. Chen, J.N., Sun, S., He, J., Torr, P.H., Yuille, A., Bai, S.: Transmix: Attend to mix for vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12135–12144 (2022)
25. Zhang, L., Deng, Z., Kawaguchi, K., Ghorbani, A., Zou, J.: How does mixup help with robustness and generalization. In: International Conference on Learning Representations (ICLR) (2021)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In:Proceedings of the IEEE conference on computer vision and pattern recognition.pp. 770–778 (2016)