Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.067024

ARTICLE

Mitigating Adversarial Attack through Randomization Techniques and Image Smoothing

Hyeong-Gyeong Kim[1], **Sang-Min Choi**[2], **Hyeon Seo**[2] **and Suwon Lee**[2,*]

[1]Artificial Intelligence Research and Development Team, Nsquare, Jinju-si, 52828, Republic of Korea
[2]Department of Computer Science and Engineering, Gyeongsang National University, Jinju-si, 52828, Republic of Korea
*Corresponding Author: Suwon Lee. Email: leesuwon@gnu.ac.kr

Received: 23 April 2025; Accepted: 12 June 2025; Published: 30 July 2025

ABSTRACT: Adversarial attacks pose a significant threat to artificial intelligence systems by exposing them to vulnerabilities in deep learning models. Existing defense mechanisms often suffer drawbacks, such as the need for model retraining, significant inference time overhead, and limited effectiveness against specific attack types. Achieving perfect defense against adversarial attacks remains elusive, emphasizing the importance of mitigation strategies. In this study, we propose a defense mechanism that applies random cropping and Gaussian filtering to input images to mitigate the impact of adversarial attacks. First, the image was randomly cropped to vary its dimensions and then placed at the center of a fixed 299 × 299 space, with the remaining areas filled with zero padding. Subsequently, Gaussian filtering with a 7 × 7 kernel and a standard deviation of two was applied using a convolution operation. Finally, the smoothed image was fed into the classification model. The proposed defense method consistently appeared in the upper-right region across all attack scenarios, demonstrating its ability to preserve classification performance on clean images while significantly mitigating adversarial attacks. This visualization confirms that the proposed method is effective and reliable for defending against adversarial perturbations. Moreover, the proposed method incurs minimal computational overhead, making it suitable for real-time applications. Furthermore, owing to its model-agnostic nature, the proposed method can be easily incorporated into various neural network architectures, serving as a fundamental module for adversarial defense strategies.

KEYWORDS: Adversarial attacks; deep learning; artificial intelligence systems; random cropping; Gaussian filtering; image smoothing

1 Introduction

Convolutional neural networks (CNNs) have demonstrated human-like or superior performance in various computer vision tasks, including object recognition, action recognition, and pose estimation [1–4]. However, CNNs and other deep-learning models are highly vulnerable to adversarial attacks [5–7].

Adversarial attacks, first identified by Szegedy et al. [8], are illustrated in Fig. 1. These attacks introduce imperceptible perturbations into the input images, generating adversarial examples that can deceive well-trained models into misclassifications. Adversarial attacks can be categorized into white- [9–11] and black-box attacks [12–14] based on the attacker's level of knowledge. Both types involve adding visually indistinguishable perturbations that cause the model to misclassify the input. In white-box attacks, an adversary has full access to the target model, including its architecture, gradients, and parameters [15]. Conversely, black-box attacks occur when the attacker has no knowledge of the model's internal structure

and can only query the input-output relationship [16]. Real-world scenarios of adversarial attacks include modifying stop signs to deceive autonomous vehicles or placing adversarial patches on objects to render them unrecognizable by artificial intelligence (AI) systems [17,18]. The rapid evolution of adversarial attack techniques poses a significant threat to AI-driven applications, particularly in computer vision and speech recognition, necessitating the development of robust defense mechanisms.

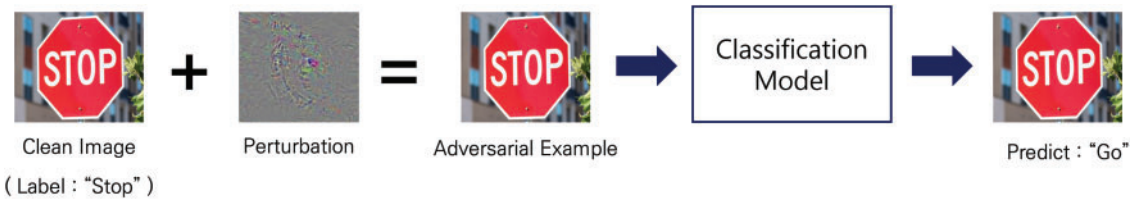

Figure 1: Example of an adversarial attack

Numerous defense strategies have been proposed to counter adversarial attacks. These strategies can be broadly classified into training- [11,19] and test-phase defenses [20–23].

One prominent training-phase defense is adversarial training [11], which incorporates adversarial examples into a training dataset to enhance the robustness of the model. Many studies have demonstrated the effectiveness of adversarial training in improving model resilience against adversarial attacks [24–26]. However, adversarial training has significant drawbacks, such as the need for model retraining whenever a new attack method is introduced, which makes it computationally expensive and time-consuming [27,28].

Contrarily, test-phase defenses do not require model retraining but, instead, focus on preprocessing input images to mitigate adversarial perturbations before feeding them into the model. One approach is adversarial verification, which applies transformations to the input images, such as flipping or cropping [21,22], or uses generative models to remove adversarial noise and reconstruct clean images [23,29]. Although adversarial purification methods eliminate the need for retraining, they are susceptible to increased inference time overhead and potential degradation in classification accuracy on clean images [27].

To address this issue, we propose a combined approach of random cropping and Gaussian filtering that focuses on image transformation during testing. The main contributions of this study are summarized as follows:

1) The proposed defense method contributes to mitigating adversarial perturbations in commonly used attack settings. While not a comprehensive solution, it demonstrates measurable improvements in robustness under standard evaluation conditions.
2) The proposed method preserves the classification performance of clean images as the applied transformations align with common data augmentation techniques, minimizing accuracy degradation.
3) This method introduces minimal inference overhead because random cropping and smoothing require low computational costs, ensuring efficient real-time deployment.
4) This method is conceptually adaptable to various network architectures and functions as a plug-and-play module, although our experiments focused on CNN-based models.

2 Related Works

2.1 Adversarial Attack Methods

Adversarial attacks introduce imperceptible perturbations to input images, causing deep learning models to misclassify them. Szegedy et al. [8] first identified this vulnerability, in which perturbations

generated adversarial examples (Fig. 1). Based on the attacker's access to model information, adversarial attacks are classified as white- and black-box attacks (Fig. 2).

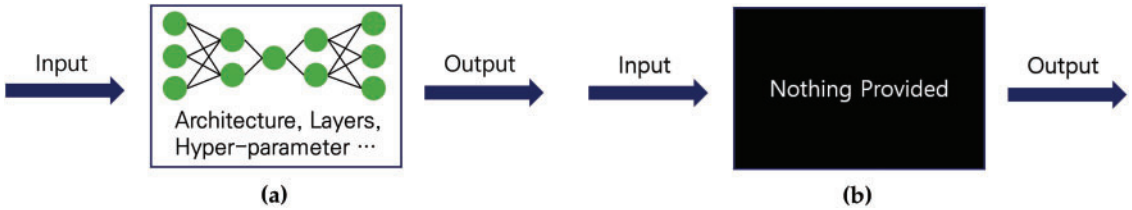

Figure 2: (**a**) White-box attack and (**b**) Black-box attack

2.1.1 White-Box Attacks

White-box attacks assume full knowledge of a model, including its architecture, parameters, and gradients [15]. The fast gradient sign method (FGSM) [9] perturbs images using a gradient sign; however, its effectiveness is limited because of its single-step nature. The projected gradient descent (PGD) [10] refines this approach by iterating multiple times. Carlini and Wagner (CW) [11] further optimized perturbations to maximize attack success while minimizing distortion [30–32]. The spatially transformed adversarial attack (stAdv) [33] deviates from pixel-level perturbations by applying spatial transformations such as smooth geometric warping. These perturbations are often visually imperceptible but result in high attack success by subtly altering object structure and positioning. The mask-guided noise restriction (MGNR) attack [34] constrains perturbations to high-saliency regions identified by binary masks, making the noise both imperceptible and highly targeted. MGNR can bypass defenses by focusing its perturbation budget on regions most influential to the model's decision.

2.1.2 Black-Box Attacks

Black-box attacks assume no access to the model's internal details and rely only on input-output queries. Transfer attacks [12] use adversarial examples crafted on a surrogate model to deceive a black box target. Query-based methods estimate the gradients using repeated queries [12]. A one-pixel attack [13] alters a single pixel to cause a misclassification, whereas a square attack [14] applies random perturbations in a block-based manner.

Although traditional black-box attacks assume access to model outputs and often involve numerous attack iterations, recent research has demonstrated more practical and efficient black-box attack methods. Park et al. introduced a similar color attack that restricted the number of attack attempts to 100 and ensured that the modifications remained imperceptible to human observers [35]. Their results showed that even with such constraints, adversarial examples can successfully deceive both deep learning models and human vision, underscoring the importance of robust defense mechanisms.

2.2 Adversarial Defense Methods

Strategies against adversarial attacks can be broadly categorized into training and test phase defenses.

2.2.1 Training-Phase Defenses

Adversarial training [9] is a widely used defense strategy that enhances model robustness by incorporating adversarial examples into the training datasets. Studies have demonstrated its effectiveness in improving resilience to various types of attacks [24–26]. However, adversarial training has some notable drawbacks. As

models should be retrained whenever a new attack method emerges, they incur high computational costs and require significant training time [27,28]. Furthermore, while adversarial training strengthens defenses against known attacks, its effectiveness against adaptive or unseen attacks remains limited [25,26]. Despite these challenges, adversarial training continues to serve as a fundamental approach in adversarial defense research, with ongoing efforts to improve its efficiency and generalizability.

2.2.2 Test-Phase Defenses

Test-phase defenses mitigate adversarial effects without modifying the model parameters and are mainly divided into generative-based purification and image transformation-based purification. Generative-based purification methods such as DiffPure [23] and DensePure [29] use diffusion models to denoise adversarial inputs. Although highly effective, these approaches incur high computational costs, limiting their real-time applicability [23,27]. Image-transformation-based purification combines input images to reduce adversarial noise with minimal computation. JPEG compression [36], random transformations [37], and spatial modifications, such as resizing and cropping [21,22] have been explored. Although computationally efficient, their defense performance is generally weaker than that of adversarial training or generative purification.

2.3 Summary of Related Works

Adversarial training provides strong robustness but requires costly retraining. Generative purification effectively removes adversarial noise but is computationally expensive. Image transformation-based defenses require no retraining and inevitably incur minimal inference overhead; however, their effectiveness is often limited. Given these tradeoffs, this study focuses on an efficient image transformation approach that balances robustness, inference speed, and broad applicability.

3 Proposed Method

In this section, we introduce our defense mechanism, which mitigates adversarial effects by combining random cropping and image smoothing using Gaussian filtering. As illustrated in Fig. 3, the proposed method comprises two key steps: (1) random cropping with zero padding and (2) image smoothing using a Gaussian filter.

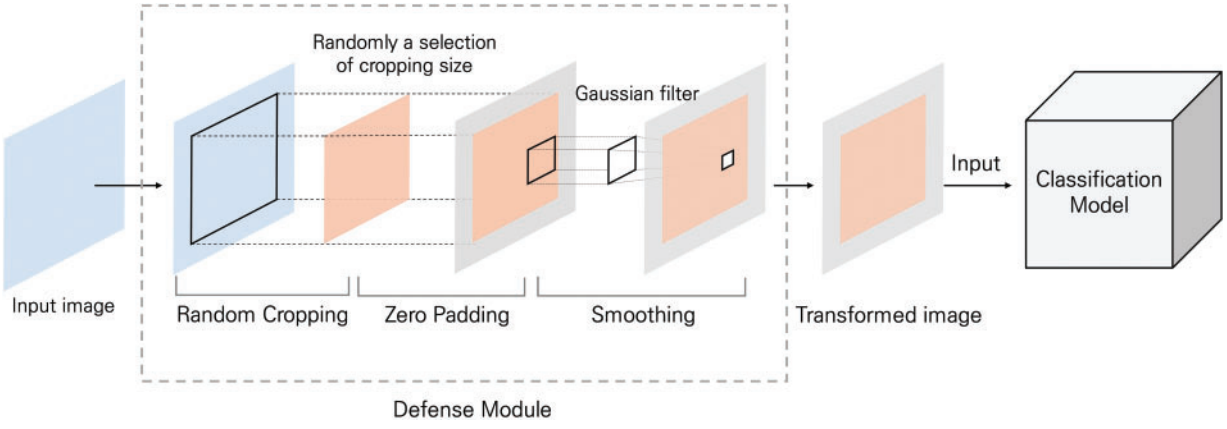

Figure 3: Overview of the proposed method

3.1 Random Cropping and Zero Padding

Random cropping alters the structure of adversarial perturbations by randomly resizing the input images. Unlike fixed-size cropping, the proposed method dynamically selects a crop size between 240 and 270 pixels at each inference step. This variability enhances the robustness against iterative attacks, such as CW and square attacks, which optimize adversarial perturbations through multiple iterations.

Once cropped, the image is resized to a fixed input dimension (299×299) using zero padding. The padding ensures that the modified image maintains a consistent shape while reducing boundary distortions that may arise during subsequent transformations.

The random cropping remains effective when applied to adversarial examples after attack generation due to its spatial disruption capability. Adversarial perturbations are typically optimized under the assumption of a fixed spatial alignment between the perturbation and the classifier's receptive fields. Randomly shifting or resizing the input invalidates this assumption by misaligning the perturbation with the model's feature extraction layers. This process disrupts the adversary's gradient optimization and diminishes the impact of location-sensitive perturbations, especially for attacks relying on structured optimization such as CW or square attacks.

3.2 Image Smoothing

After random cropping and padding, Gaussian filtering is applied to smooth the image and reduce the effectiveness of adversarial perturbations. Image smoothing modifies pixel values based on neighboring pixels, thereby disrupting adversarial noise patterns.

Among the various smoothing techniques, including mean, median, and bilateral filtering [38], Gaussian filtering was selected owing to its superior adversarial mitigation performance. Gaussian filters assign higher weights to closer pixels following a Gaussian distribution, effectively suppressing noise while preserving the image structure.

The effectiveness of Gaussian filtering as a post-processing defense stems from the frequency characteristics of adversarial noise. Most adversarial perturbations are high-frequency signals that exploit the model's sensitivity to minor pixel-level changes. Gaussian filtering acts as a low-pass filter, attenuating these high-frequency components while preserving the semantic content of the image. Even when applied after the adversarial example has been crafted, this operation significantly weakens the adversarial signal embedded in the image, reducing its ability to mislead the classifier.

3.3 Defense Process with Random Cropping and Gaussian Filtering

The overall process of the proposed defense mechanism is as follows:

1. The input image is received for classification.
2. A random integer between 240 and 270 is selected.
3. The image is cropped to the randomly selected size.
4. The cropped image is centered in a 299×299 space.
5. Zero padding is applied to fill the remaining space.
6. Gaussian filtering with a 7×7 kernel and a standard deviation of two is applied.
7. The transformed image is passed to the classification model.

Instead of feeding the input image directly into the classification model, the image is randomly cropped to vary its dimensions. The cropped image is then placed at the center of a fixed 299×299 space, with the remaining areas filled with zero padding. Subsequently, Gaussian filtering with a 7×7 kernel and a standard deviation of two is applied using a convolution operation.

Finally, the smoothed image is input into the classification model. This transformation effectively alters the structure of adversarial perturbations, reducing their impact while maintaining classification accuracy on clean images. Fig. 4 shows an example of the original image (a) and its transformed version (b) using the proposed defense mechanism.

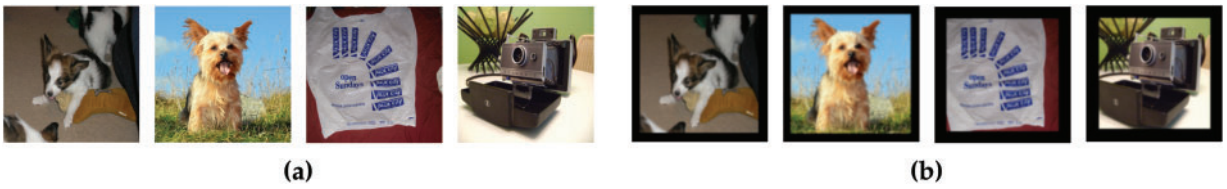

Figure 4: Examples of images converted by the proposed method: (**a**) original image and (**b**) result image

4 Experimental Result

4.1 Experimental Setup

We tested 10,000 ImageNet validation images [39] correctly classified using each model. The hardware used is listed in Table 1. ResNet101 [40] served as the main target for defense evaluation, and Inception v3 [41] and VGG16 [42] were used to examine the generality.

Table 1: Experimental environment

Component	Specification
CPU	Intel i7-12700k
GPU	GeForce RTX 3060
RAM	DDR5 64 GB
OS	Windows 11

4.2 Adversarial Example Generation

To generate adversarial examples, we employed three attack methods: white-box single-step FGSM, white-box iterative CW, black-box Square, stAdv, and MGNR attacks. The FGSM perturbations became more severe as the perturbation factor increased. We conducted attacks with FGSM using ϵ values of $\{0.01, 0.02, 0.05\}$. For CW and square attacks, we set a maximum of 1000 iterations but allowed early stopping if the loss value remained unchanged. For the stAdv attack, we applied spatial transformations by optimizing a flow field via L-BFGS with 500 iterations. The optimization minimizes both adversarial loss and total variation of the flow field, with loss weights $\lambda_{tv} = 50$ and $\lambda_{adv} = 0.05$. For the MGNR attack, adversarial perturbations were applied only to semantically important regions defined by pre-computed masks. We used a PGD-based method with 100 iterations, a step size of $\alpha = 2/255$, and a maximum perturbation of $\varepsilon = 16/255$, enforcing perturbation only within the masked areas.

Using 10,000 correctly classified images from the dataset, we generated 10,000 adversarial examples for each attack method, which resulted in 70,000 adversarial images.

Fig. 5 illustrates examples of adversarial images generated using different attack methods. FGSM, which applies perturbations in a single step, spreads adversarial patterns across the entire image, with increasing ε values making the perturbations more pronounced. The CW attack, an iterative method, introduces nearly imperceptible perturbations in specific regions rather than across an entire image. Square attacks generate

adversarial patterns in the form of rectangular noise patches. The stAdv attack distorts spatial structure by applying smooth geometric transformations, resulting in perceptually similar images with spatially shifted features.

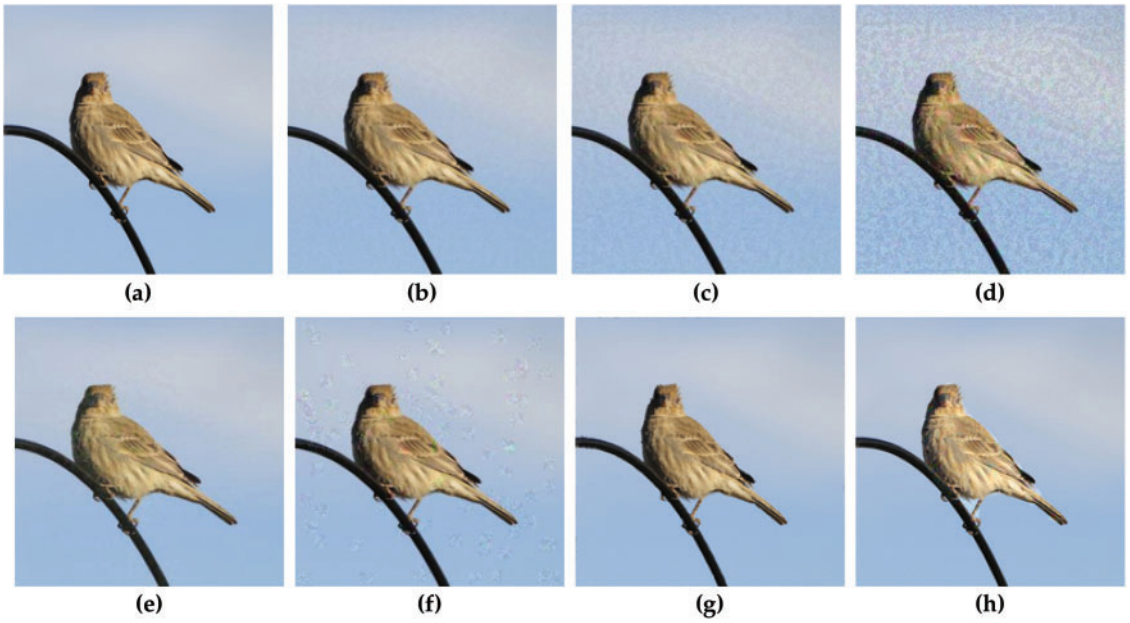

Figure 5: Examples of the generated adversarial images: (**a**) clean image, (**b**) FGSM (ε = 0.01), (**c**) FGSM (ε = 0.02), (**d**) FGSM (ε = 0.05), (**e**) CW, (**f**) Square, (**g**) stAdv, and (**h**) MGNR

The MGNR attack confines perturbations to semantically important regions using binary masks, producing localized distortions that remain largely imperceptible in the unmasked background.

4.3 Evaluation of Adversarial Defense Mechanisms

We compared the proposed defense mechanisms (random cropping and Gaussian filtering) with existing adversarial defense methods, including random resizing with padding [21] and random flipping with cropping [22]. Moreover, we evaluate various image-smoothing techniques. Each defense method applies a pre-processing transformation to the input image before feeding it into the classifier. Fig. 6 illustrates the images transformed using different defense mechanisms.

Figure 6: (Continued)

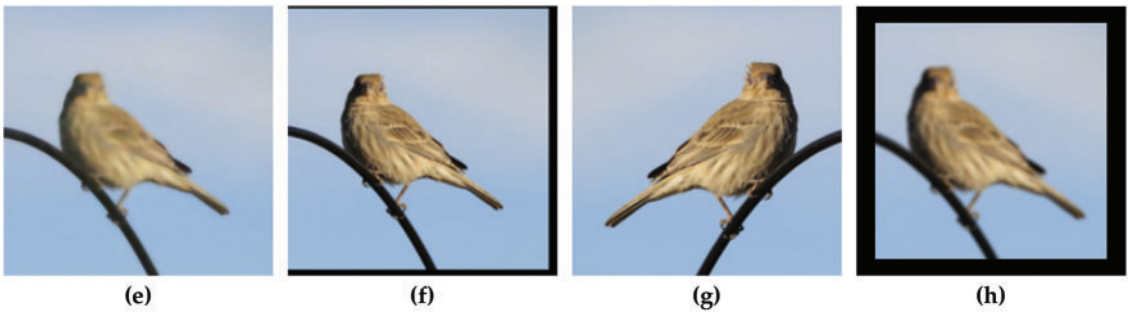

Figure 6: Examples of an image transformed using different defense techniques: (**a**) clean image, (**b**) mean filtering, (**c**) median filtering, (**d**) gaussian filtering, (**e**) bilateral filtering, (**f**) random resizing and padding, (**g**) flip and crop, (**h**) proposed method

4.4 Performance Degradation on Clean Images

A key evaluation criterion for adversarial defense is the effect on clean images. Considering a defense algorithm is applied to all input images, regardless of whether they are adversarial, excessive performance degradation on clean images can undermine model reliability.

We measured the classification accuracy of ResNet101 for clean images using various defense methods, with the results summarized in Table 2. The baseline ResNet101 model achieved 100% accuracy as the dataset was curated to exclude misclassified samples. Most defense methods resulted in a minor accuracy drop of approximately 2%, whereas median and mean filtering exhibited the highest accuracy drop of 10.1%. The flip-and-crop strategy was the best-performing method in terms of preserving accuracy, with accuracy reduced by only 0.62%.

Table 2: Accuracy of clean images for different defense methods

Defense method	Accuracy (%)
ResNet101 (No defense)	100
ResNet101 + Flip and crop	99.38
ResNet101 + Random resizing and padding	97.94
ResNet101 + Median filter	89.90
ResNet101 + Mean filter	89.90
ResNet101 + Gaussian filter	98.19
ResNet101 + Bilateral filter	90.30
ResNet101 + Proposed method	98.18

4.5 Inference Time Evaluation

An adversarial defense mechanism does not impose an excessive inference overhead. We measured the per-image inference time for ResNet101 with and without each defense method, with the results summarized in Table 3. The baseline model, without any defense, required 21.20 ms per image. Most defense methods, including the proposed approach, had a negligible impact on the inference time, except for median and bilateral filtering, which significantly increased the computational overhead.

Table 3: Inference time per image for different defense methods

Defense method	Inference time (ms)
ResNet101 (No defense)	21.20
ResNet101 + Flip and crop	21.23
ResNet101 + Random resizing and padding	22.70
ResNet101 + Median filter	88.11
ResNet101 + Mean filter	21.23
ResNet101 + Gaussian filter	21.25
ResNet101 + Bilateral filter	27.43
ResNet101 + Proposed method	21.50

4.6 Classification Accuracy on Adversarial Examples

To assess the effectiveness of each defense mechanism, we applied ResNet101 to the adversarial examples generated in Section 4.2. Median and mean filtering were excluded because of significant accuracy degradation and inference time overhead.

Table 4 presents the classification accuracy of adversarial examples under various attack scenarios. Higher accuracy indicates greater mitigation of adversarial effects. All defense methods showed improvements over the baseline model without defense.

Table 4: Classification accuracy on adversarial examples for different defense methods. The highest accuracy for each adversarial example is highlighted in bold

Defense method	Accuracy (%)						
	FGSM			CW	Square	stAdv	MGNR
	(ε = 0.01)	(ε = 0.02)	(ε = 0.05)				
ResNet101 + No defense	36.10	23.30	25.57	33.14	61.40	58.27	72.12
ResNet101 + Flip and crop	61.68	45.92	44.52	53.95	75.44	85.00	83.62
ResNet101 + Random resizing and padding	72.89	56.11	48.26	68.53	80.91	84.04	84.04
ResNet101 + Gaussian filter	80.28	63.81	59.68	67.79	84.55	**85.19**	86.35
ResNet101 + Proposed method	**87.78**	**81.34**	**70.93**	**77.65**	**87.78**	84.54	**87.31**

Although the highest accuracy does not occur for every single attack type, the proposed method consistently achieved high performance across all evaluated threats, including FGSM, CW, Square, stAdv, and MGNR attacks. For FGSM attacks with ε = 0.01, the accuracy improved from 36.10% to 87.78%. Similar mitigation was observed for CW, Square, and MGNR attacks, where the proposed method ranked among the top performers while maintaining stable performance across all threat models.

These results suggest that the proposed method provides balanced and effective mitigation across diverse adversarial scenarios, demonstrating its suitability as a general-purpose, test-time defense mechanism.

Furthermore, an ablation comparison among random cropping only (implemented via random resizing and padding), Gaussian filtering only, and their combination reveals clear synergy. Gaussian filtering alone contributes more to adversarial mitigation than cropping alone. However, combining both techniques leads

to the highest overall robustness, indicating that the two transformations complement each other in reducing adversarial impact.

4.7 Evaluation of Classification Accuracy by Parameter Selection

Experiments were conducted using two key parameters–random cropping size and Gaussian filter parameters–to determine the optimal parameters for the proposed defense method. The classification accuracy of the clean and adversarial examples was evaluated for each setting, with the optimal values selected based on the results.

4.7.1 Evaluation of Classification Accuracy by Random Cropping Size

The proposed defense method uses random cropping within a predefined integer range to modify the input image size. Experiments were conducted without Gaussian filtering to analyze the impact of the cropping size on the classification performance. Image sizes were set to {200, 210, 220, 230, 240, 250, 260, 270, 280, 290, and 299}, and classification accuracy was measured for both clean images and adversarial examples generated by FGSM ($\varepsilon = 0.01$).

The results shown in Fig. 7 indicate that the classification accuracy on adversarial examples was the highest when images were cropped to sizes between 240 × 240 and 270 × 270. Furthermore, minimal degradation in accuracy was observed for clean images with these cropping sizes. Based on these findings, the random cropping size in the proposed method was set to a random integer between 240 and 270 pixels to effectively mimic adversarial attacks while maintaining a high classification accuracy.

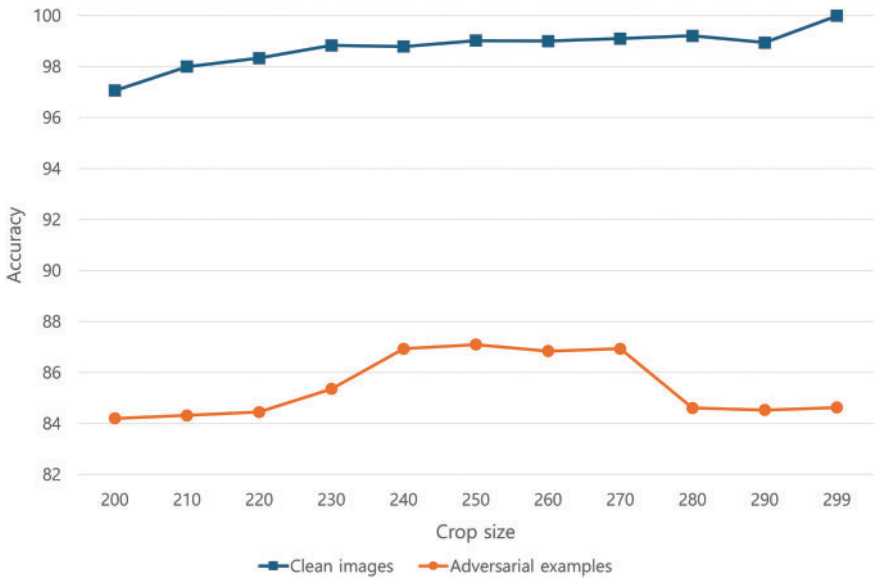

Figure 7: Classification accuracy by random cropping size

4.7.2 Evaluation of Classification Accuracy by Gaussian Filtering Parameters

Gaussian filtering involves two critical parameters: the kernel size and standard deviation (σ). To analyze their impact, experiments were conducted without applying random cropping by varying the kernel sizes {3, 5, 7, and 9} and the standard deviation {1.0, 2.0, 3.0, 4.0, and 5.0}. The classification accuracy was evaluated separately for clean images and adversarial examples.

The results, illustrated in Fig. 8, show that increasing the kernel size generally improves robustness against adversarial examples, with the best adversarial mitigation observed at kernel sizes seven and nine, with $\sigma = 2$. However, using a kernel size of nine significantly reduced the accuracy of clean images, suggesting that excessive smoothing distorted the image features.

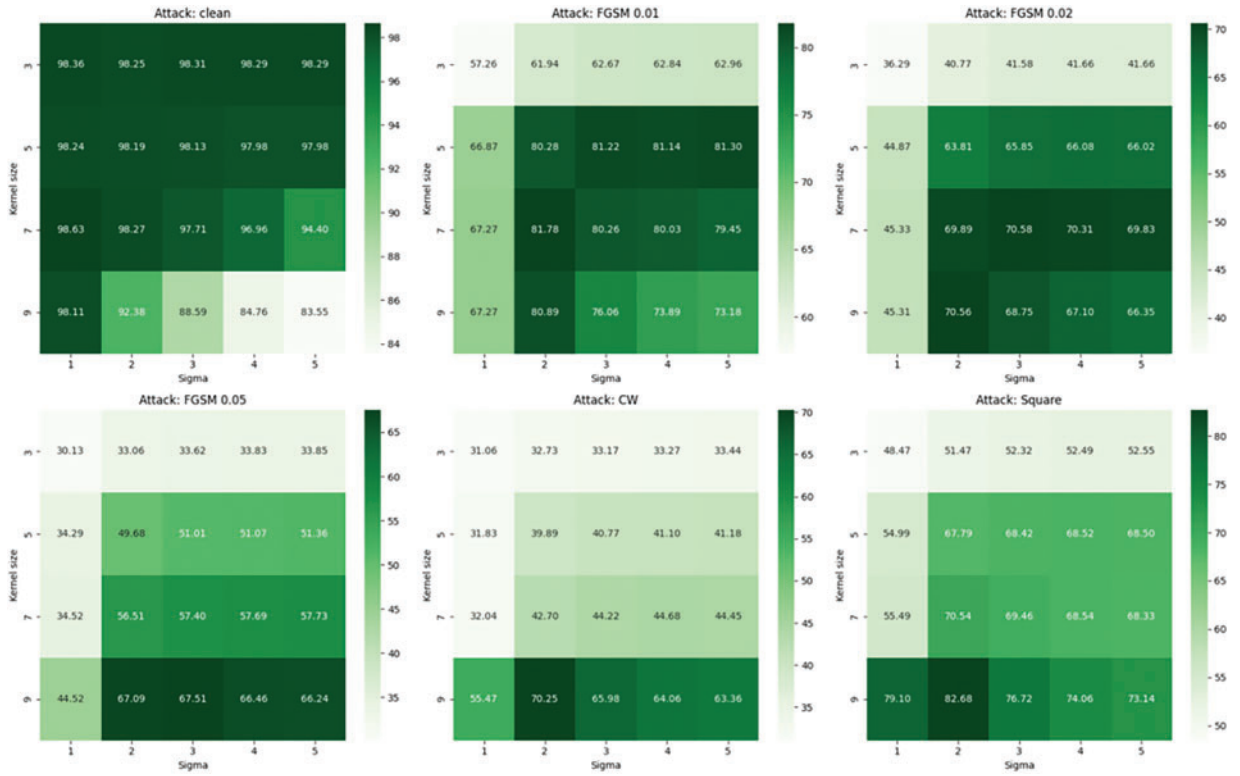

Figure 8: Classification accuracy by Gaussian filter kernel size and standard deviation

Thus, to maintain high accuracy on both clean and adversarial examples, the optimal Gaussian filter parameters were selected as follows:

- Kernel size: 7×7
- Standard deviation (σ): 2.0

4.8 Generalization across Different Models

To evaluate the generalizability of the proposed method, we conducted additional experiments using different CNN-based model architectures. Specifically, the method was tested on pre-trained **Inception V3** and **VGG16** without additional training or fine-tuning.

Since the adversarial examples used in previous experiments were generated against ResNet101, new adversarial examples were generated for both Inception V3 and VGG16. For each of the seven attack types, we created **5000 adversarial examples per model**, resulting in a total of **70,000 adversarial examples**.

Classification accuracy was measured with and without the defense method to assess its effectiveness. Results presented in Table 5 show that applying the proposed method led to substantial accuracy improvements for both Inception V3 and VGG16. These findings suggest that the defense may generalize well across multiple **CNN-based classifiers** and effectively mitigates adversarial attacks.

Table 5: Classification accuracy of different models with the proposed defense method

Model	Accuracy (%)						
	FGSM ($\varepsilon = 0.01$)	FGSM ($\varepsilon = 0.02$)	FGSM ($\varepsilon = 0.05$)	CW	Square	stAdv	MGNR
Inception V3 (No defense)	36.40	27.82	26.14	30.17	63.22	62.10	71.51
Inception V3 + Proposed method	**86.32**	**80.75**	**71.45**	**75.11**	**87.93**	**85.00**	**84.45**
VGG16 (No defense)	34.38	22.40	23.62	35.94	61.07	53.46	68.15
VGG16 + Proposed method	**86.13**	**78.75**	**67.36**	**71.53**	**82.61**	**82.52**	**77.02**

4.9 Visualization of Defense Performance

To comprehensively compare the effectiveness of the different defense methods, we visualized the classification accuracy results, as shown in Fig. 9. Each subfigure represents the classification accuracy on adversarial examples generated by different attack methods: (a) FGSM ($\varepsilon = 0.01$), (b) FGSM ($\varepsilon = 0.02$), (c) FGSM ($\varepsilon = 0.05$), (d) CW, and (e) square attacks.

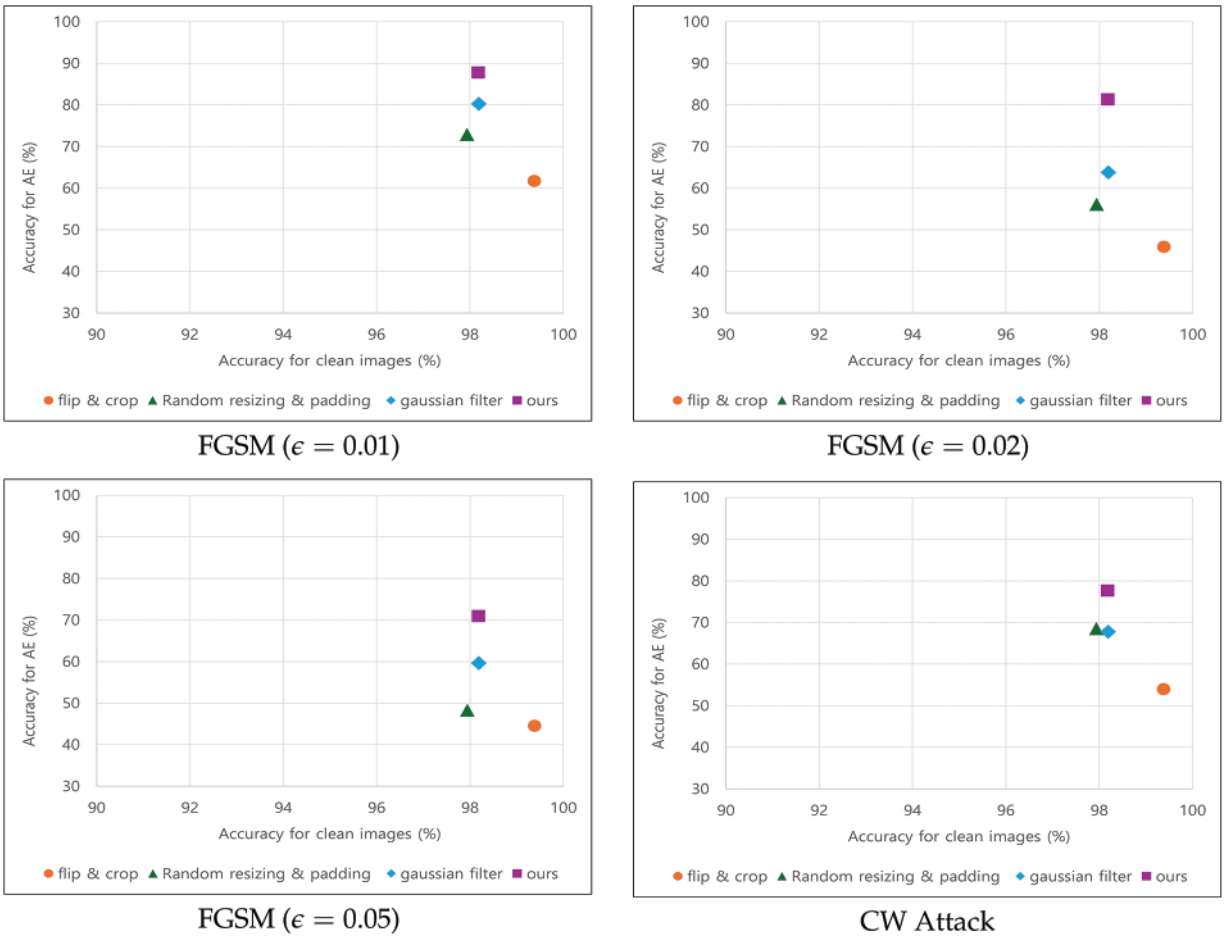

Figure 9: (Continued)

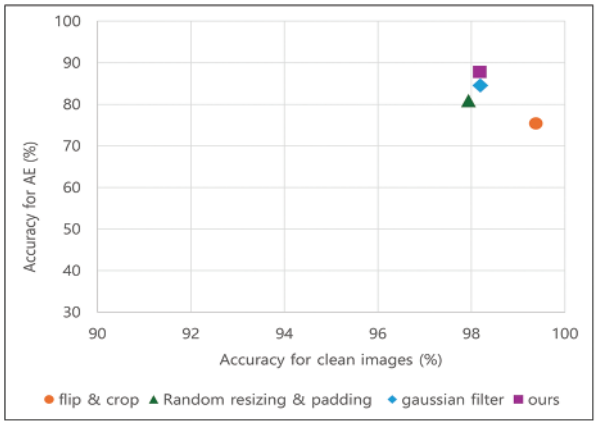

Figure 9: Classification accuracy visualization for different defense methods across adversarial attacks. *AE: Adversarial Examples

In these plots, the x-axis represents the classification accuracy on clean images, whereas the y-axis represents the classification accuracy on adversarial examples. A defense method positioned toward the top-right corner of the graph indicates superior performance because it maintains a high classification accuracy on both clean and adversarial images while effectively mitigating adversarial attacks.

As shown in Fig. 9, the proposed defense method consistently appeared in the upper-right region across all attack scenarios, demonstrating its ability to preserve classification performance on clean images while significantly mitigating adversarial attacks. This visualization confirms that the proposed method is effective and reliable for defending against adversarial perturbations.

5 Discussion

While the proposed method demonstrates strong mitigation against various adversarial attacks, including gradient-based (FGSM, CW), query-based (Square), spatial (stAdv), and semantically guided (MGNR) perturbations, several limitations remain.

First, although our approach operates entirely at the input level and does not rely on model-specific gradients or architecture-dependent assumptions, our empirical evaluation was limited to convolutional neural networks (CNNs), including ResNet101, Inception V3, and VGG16. Architectures such as Vision Transformers (ViTs) differ fundamentally in their handling of spatial information due to their reliance on positional embeddings and lack of translation equivariance; thus, further studies are needed to assess whether the proposed transformations–particularly random cropping-interact differently with such models.

Second, the evaluation focused on standardized image classification datasets, with ImageNet serving as the primary benchmark. While ImageNet provides high-resolution and semantically diverse content, its scale and complexity may not capture domain-specific challenges present in medical or satellite imagery. Moreover, smoothing-based defenses such as Gaussian filtering are less effective on low-resolution datasets such as CIFAR-10, where the risk of semantic information loss increases with even modest filtering. Despite this, prior research suggests that defense methods that perform well on ImageNet often generalize to other domains, although the reverse is not always true. Nonetheless, applying the proposed method to additional domains remains an important future direction.

Third, our study does not evaluate adaptive attacks in which the attacker is aware of the defense and actively seeks to circumvent it. Although our method effectively mitigates several strong attacks, including those localized in semantically important regions (e.g., MGNR), a more rigorous evaluation against adaptive adversaries–such as those generated via Expectation over Transformation (EOT) or smoothing-aware optimization–would further validate the robustness of the defense under stronger threat models.

Fourth, while the core of our method leverages established techniques (random cropping and Gaussian filtering), our contribution lies in demonstrating their synergistic combination and practical utility as a test-time, plug-and-play module. We have provided theoretical reasoning (Sections 3.1 and 3.2) to support the effectiveness of these transformations in disrupting adversarial gradients and attenuating high-frequency noise. However, further exploration into their effects on model calibration, prediction confidence, and loss landscapes could yield deeper insights.

Finally, we deliberately focused our comparisons on test-time, preprocessing-based defenses to align with the computational constraints and design philosophy of our method. While certified defenses, randomized smoothing, and diffusion-based purification approaches offer promising robustness guarantees, they often require retraining or incur high inference-time costs. A more comprehensive comparison across different defense paradigms remains a valuable direction for future work.

Future work will focus on (i) extending evaluation to non-CNN architectures and domain-specific datasets, (ii) testing the defense against fully adaptive adversaries, and (iii) comparing our lightweight approach with certified or generative defenses under a unified cost–robustness framework.

6 Conclusion

In this study, we proposed a lightweight, test-time defense method that integrates random cropping and Gaussian filtering to suppress adversarial perturbations. Unlike training-based defenses or generative purification techniques, the proposed method operates entirely at the input level without requiring model retraining or architectural modifications, incurring minimal inference-time overhead–making it suitable for real-time deployment.

Extensive experiments demonstrated that the proposed method consistently improves classification accuracy against a wide range of adversarial threats, including gradient-based (FGSM, CW), query-based (Square), spatial (stAdv), and semantically guided (MGNR) attacks. Furthermore, evaluation on ResNet101, Inception V3, and VGG16 confirmed that the performance of the proposed method is robust across CNN-based architectures, supporting its practical applicability and its potential for architectural generalization within the CNN domain.

While the results are promising, the current scope is limited to CNN architectures and standard classification benchmarks. Future work will explore the applicability of the proposed method to non-CNN models such as ViTs, evaluate its robustness under fully adaptive attack settings, and compare its efficiency and effectiveness against certified and generative defenses within a unified cost–robustness framework.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Glocal University 30 Project Fund of Gyeongsang National University in 2025.

Author Contributions: Study conception and design: Hyeong-Gyeong Kim, Suwon Lee; data collection: Sang-Min Choi, Hyeon Seo; analysis and interpretation of results: Hyeong-Gyeong Kim, Suwon Lee; draft manuscript preparation: Hyeong-Gyeong Kim, Sang-Min Choi, Hyeon Seo; revision of the manuscript: Hyeong-Gyeong Kim, Suwon Lee. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials used in this study are currently part of an ongoing project and cannot be publicly released at this time. Access to the data may be considered upon reasonable request after the completion of the project.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Greenspan H, Van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging. 2016;35(5):1153–9. doi:10.1109/tmi.2016.2553401.
2. Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell. 2017;40(6):1510–7. doi:10.1109/tpami.2017.2712608.
3. Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI, USA; 2017. p. 7291–9.
4. Silva SH, Najafirad P. Opportunities and challenges in deep learning adversarial robustness: a survey. arXiv:2007.00753. 2020.
5. Huang S, Papernot N, Goodfellow I, Duan Y, Abbeel P. Adversarial attacks on neural network policies. arXiv:1702.02284. 2017.
6. Yan Z, Guo Y, Zhang C. Deep defense: training DNNs with improved adversarial robustness. In: Advances in neural information processing system. Montreal, QC, Canada; 2018. Vol. 31.
7. Deng Y, Zheng X, Zhang T, Chen C, Lou G, Kim M. An analysis of adversarial attacks and defenses on autonomous driving models. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom); Austin, TX, USA; 2020. p. 1–10.
8. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR); Banff, AB, Canada; 2014.
9. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR); San Diego, CA, USA; 2015.
10. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR); Vancouver, BC, Canada; 2018.
11. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy (SP); San Jose, CA, USA; 2017. p. 39–57.
12. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the ACM Asia Conference on Computer and Communications Security (AsiaCCS); Abu Dhabi, United Arab Emirates; 2017. p. 506–19.
13. Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput. 2019;23(5):828–41.
14. Andriushchenko M, Croce F, Flammarion N, Hein M. Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision (ECCV); Glasgow, UK; 2020. p. 484–501.
15. Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst. 2019;30(9):2805–24. doi:10.1109/tnnls.2018.2886017.
16. Costa JC, Roxo T, Proença H, Inácio PRM. How deep learning sees the world: a survey on adversarial attacks & defenses. IEEE Access. 2024;12:61113–36.
17. Duan R, Ma X, Wang Y, Bailey J, Qin AK, Yang Y. Adversarial camouflage: hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Seattle, WA, USA; 2020. p. 1000–8.

18. Thys S, Van Ranst W, Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Long Beach, CA, USA; 2019.
19. Carlini N, Wagner D. Defensive distillation is not robust to adversarial examples. arXiv:1607.04311. 2016.
20. Zhang C, Gao P. Countering adversarial examples: combining input transformation and noisy training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada; 2021. p. 102–11.
21. Xie C, Wang J, Zhang Z, Ren Z, Yuille A. Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (ICLR); Vancouver, BC, Canada; 2018.
22. Pérez JC, Alfarra M, Jeanneret G, Rueda L, Thabet A, Ghanem B, et al. Enhancing adversarial robustness via test-time transformation ensembling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); Montreal, QC, Canada; 2021. p. 81–91.
23. Nie W, Guo B, Huang Y, Xiao C, Vahdat A, Anandkumar A. Diffusion models for adversarial purification. In: International Conference on Machine Learning (ICML); Baltimore, MD, USA; 2022. p. 16805–27.
24. Zhang H, Yu Y, Jiao J, Xing E, El Ghaoui L, Jordan M. Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (ICML); Long Beach, CA, USA; 2019. p. 7472–82.
25. Wang J, Zhang H. Bilateral adversarial training: towards fast training of more robust models against adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); Long Beach, CA, USA; 2019. p. 6629–38.
26. Zhang J, Zhu J, Niu G, Han B, Sugiyama M, Kankanhalli M. Geometry-aware instance-reweighted adversarial training. In: International Conference on Learning Representations (ICLR); Vancouver, BC, Canada; 2018.
27. Choi SH, Kim HG, Choi YH. An adversarial attack type classification method using linear discriminant analysis and k-means algorithm. J Korea Inst Inf Secur Cryptol. 2021;31(6):1215–25.
28. Laykaviriyakul P, Phaisangittisagul E. Collaborative Defense-GAN for protecting adversarial attacks on classification system. Expert Syst Appl. 2023;214:118957.
29. Chen Z, Jin K, Wang J, Nie W, Liu M, Anandkumar A, et al. Densepure: understanding diffusion models towards adversarial robustness. In: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS; New Orleans, LA, USA; 2022.
30. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P); Saarbruecken, Germany; 2016. p. 372–87.
31. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. p. 2574–82.
32. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Boston, MA, USA; 2015. p. 427–36.
33. Xiao C, Li B, Zhu JY, He W, Liu M, Song D. Spatially transformed adversarial examples. In: International Conference on Learning Representations (ICLR); Vancouver, BC, Canada; 2018.
34. Duan Y, Zhou X, Zou J, Qiu J, Zhang J, Pan Z. Mask-guided noise restriction adversarial attacks for image classification. Comput Secur. 2021;100:102111.
35. Park D, Yeon S, Seo H, Buu S, Lee S. Practical adversarial attacks imperceptible to humans in visual recognition. Comput Model Eng Sci. 2025;142(3):2725–37. doi:10.32604/cmes.2025.061732.
36. Liu Z, Liu Q, Liu T, Xu N, Lin X, Wang Y, et al. Feature distillation: DNN-oriented JPEG compression against adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 860–8.
37. Raff E, Sylvester J, Forsyth S, McLean M. Barrage of random transforms for adversarially robust defense. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Long Beach, CA, USA; 2019. p. 6528–37.

38. Li P, Wang H, Yu M, Li Y. Overview of image smoothing algorithms. J Phys Conf Ser. 2021;1883(1):12–24.
39. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Miami, FL, USA; 2009.
40. 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 (CVPR); Las Vegas, NV, USA; 2016. p. 770–8.
41. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA; 2016. p. 2818–26.
42. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR); San Diego, CA, USA; 2015. p. 1–14.