Signal, Image and Video Processing (2025) 19:481
https://doi.org/10.1007/s11760-025-04038-2

ORIGINAL PAPER

Multi-resolution training improves robustness against adversarial attacks

Shuaiang Rong[1] · **Emadeldeen Hamdan**[1] · **Ahmet Enis Cetin**[1]

Received: 5 December 2024 / Revised: 19 February 2025 / Accepted: 5 March 2025 / Published online: 15 April 2025

Abstract
Deep neural networks (DNNs) have progressed rapidly in recent years and are increasingly deployed in real-world applications. They are now integral to critical tasks, such as traffic sign recognition in autonomous vehicles, where DNNs have become the primary method for handling most of the processing. However, many DNNs are known to be vulnerable to adversarial attacks-small but deliberately crafted perturbations applied to input data. Such perturbations can easily cause misclassification, posing significant risks, especially in autonomous vehicle systems. In this paper, we present a novel approach called multi-resolution training, which utilizes lower-resolution information from input images to retain essential features while partially filtering out adversarial attacks. Our method involves designing convolutional neural network (CNN) layers that apply various downsampling techniques with custom-designed filters, followed by upsampling to restore the resolution for further network processing. This approach has been tested on multiple DNNs, and results show that it effectively enhances the robustness of DNNs against adversarial attacks.

Keywords Deep learning neural network · Adversarial attack · Low-pass filter · Gaussian filter

1 Introduction

Recent research has demonstrated that DNNs are surprisingly vulnerable to adversarial examples, where small modifications to input data can result in significant incorrect predictions [1, 2]. This susceptibility arises from DNNs' inherent linearity in high-dimensional spaces and limited generalization. This vulnerability is especially concerning for traffic sign recognition systems, which rely exclusively on DNNs processing camera images without additional error correction sources.

Many research efforts have been made on attacks targeting these systems, as they are particularly susceptible and easily exploited in real-world scenarios [3]. Early research into adversarial attacks on traffic sign recognition typically involved physical methods, such as affixing stickers to traffic signs [4], replacing signs with versions that included embedded perturbations [3], or placing patches directly on camera lenses [5]. While these methods can be effective, they require manual effort and are often noticeable to drivers.

To address these limitations, researchers have shifted their focus to direct attacks on DNN inputs. One of the first developed techniques is the Fast Gradient Sign Method (FGSM), which uses the gradients of the loss function relative to the input data to create perturbations that maximize the loss and result in misclassification [2]. Subsequent developments include the Projected Gradient Descent (PGD) approach [6]. Other significant adversarial algorithms are the Jacobian-based Saliency Map Attack (JSMA) [7], Papernot's attack [8], and the Carlini and Wagner (C&W) attack [9], all contributing to the evolving adversarial machine learning. Figure 1 illustrates how FGSM attack corrupt captured images and alter the predictions of the DNN.

Despite the proliferation of these attack methods, defense and mitigation strategies, particularly for traffic sign classification tasks, have received significantly less attention. The common approach to enhancing the robustness of DNNs is adversarial training, which incorporates adversarial attack

✉ Ahmet Enis Cetin
aecyy@uic.edu

Shuaiang Rong
srong4@uic.edu

Emadeldeen Hamdan
ehamda3@uic.edu

[1] Department of Electrical and Computer Engineering, University of Illinois Chicago, 1200 West Harrison St, Chicago, IL 60607, USA

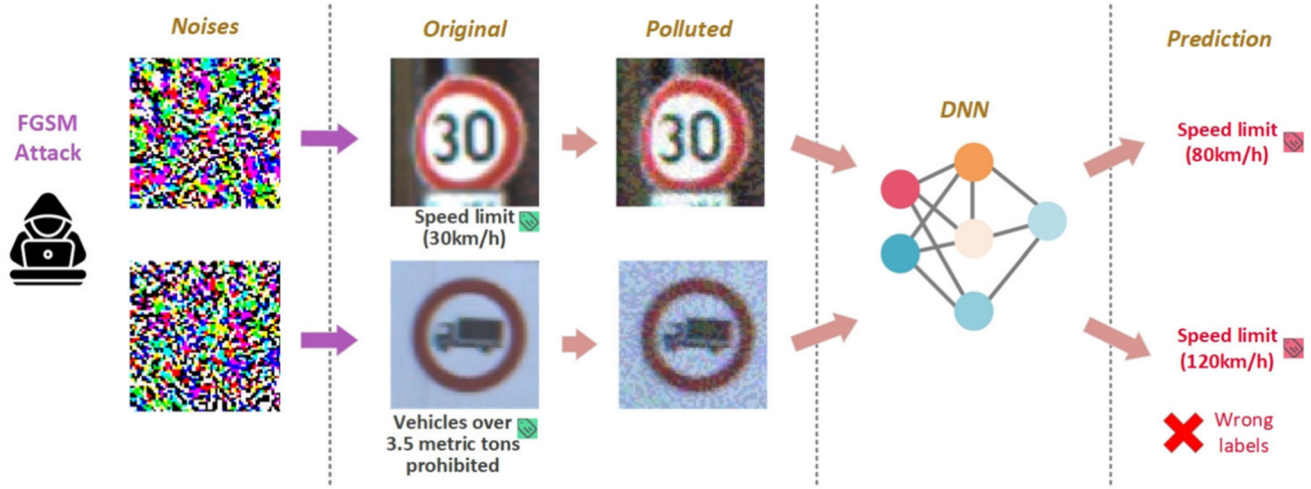


Fig. 1 DNN vulnerability against adversarial attack

samples into the training dataset [10]. Training a model with only one or a limited range of adversarial examples leaves it susceptible to other types of attacks, necessitating the inclusion of diverse adversarial examples during training. Consequently, this approach is not universal and can demand significant training effort.

2 Multi-resolution training

We propose a novel multi-resolution training approach by enhancing DNN architectures with the integration of an additional CNN block prior to the main network, as shown in Fig. 2. This block processes traffic sign images by first downsampling them to a lower resolution through decimation and then upsampling them back to the original resolution via interpolation. The block outputs either a 3-channel RGB image directly fed into the network or a 6-channel output obtained by concatenating the processed image with the original.

The CNN block utilizes layers with filters designed using various downsampling techniques, such as low-pass (LP) and Gaussian filtering. This approach effectively reduces the impact of subtle, high-frequency adversarial perturbations while preserving the essential features of traffic signs, thereby improving the robustness of DNNs against adversarial attacks.

2.1 Multi-resolution CNN block

The multi-resolution image process is implemented through a custom-designed CNN block, which includes a downsampling CNN layer with multiple filter options, followed by a standard bilinear interpolation layer for upsampling.

The first option utilizes a simple 1D low-pass filter defined as $h_{lp} = [\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$, commonly used for its simplicity and effectiveness in smoothing operations, approximating a half-band low-pass filter. The corresponding 2D filter used in the *LP_conv* layer is computed as:

$$h_{\text{lp-2D}}[n_1, n_2] = h_{\text{lp}} \cdot h_{\text{lp}}^T$$

which results in a 3×3 kernel:

$$h_{\text{lp-2D}} = \begin{bmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{bmatrix}.$$

This 2D filter effectively smooths input image data, making it ideal for downsampling in the multi-resolution framework.

The second option is a Gaussian filter, which offers more customization and smoother outputs. The normalized 1D Gaussian filter is defined as:

$$h_{\text{g}}[n] = \frac{\exp\left(-0.5\left(\frac{n}{\sigma}\right)^2\right)}{\sum_n \exp\left(-0.5\left(\frac{n}{\sigma}\right)^2\right)}$$

where n denotes the indices, represented as $(-\frac{z-1}{2}, \ldots, 0, \ldots, \frac{z-1}{2})$. z is the filter size, and σ is the standard deviation, which controls the spread of the filter. The resulting 2D Gaussian filter applied in the *Gaussian_conv* layer is computed as:

$$h_{\text{g-2D}}[n_1, n_2] = h_{\text{g}} \cdot h_{\text{g}}^T$$

By adjusting the filter size z and the standard deviation σ, the *Gaussian_conv* layer can achieve varying levels of smoothing, offering enhanced flexibility. This design

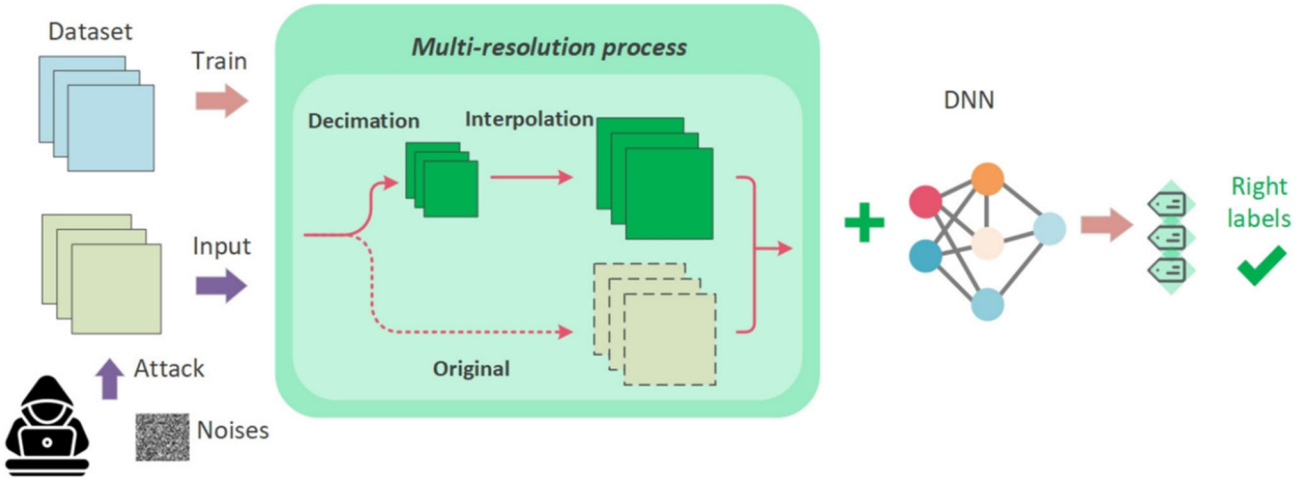


Fig. 2 Multi-resolution training process

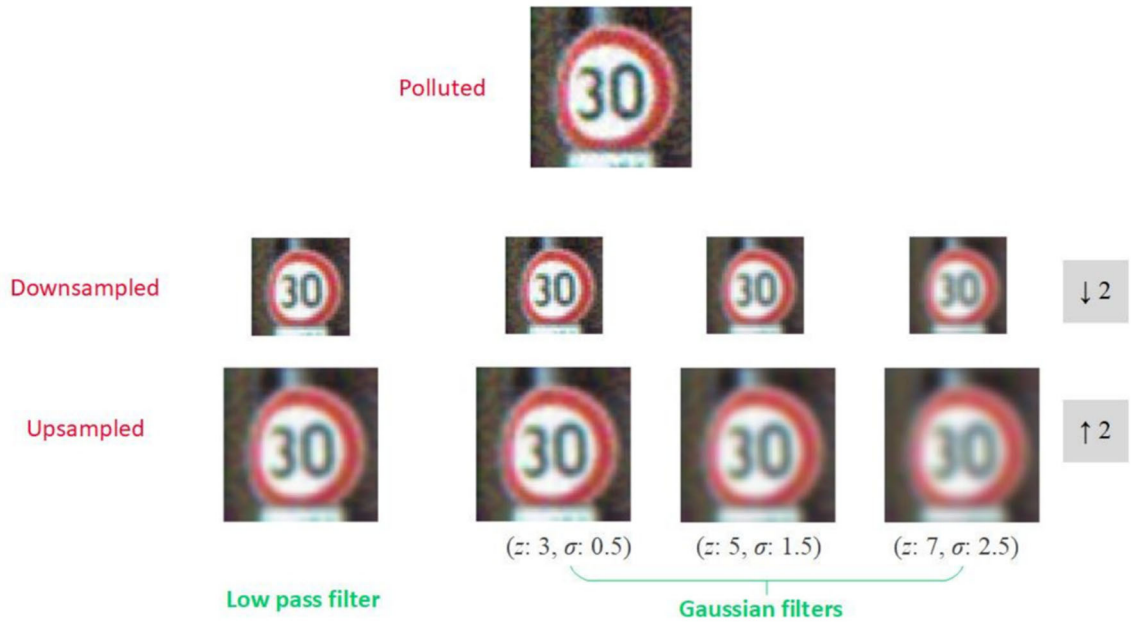


Fig. 3 Image processing using LP and Gaussian conv layer

ensures that the multi-resolution training framework adapts to different filtering needs while maintaining computational efficiency. Figure 3 presents the processing results of polluted images using the designed CNN block. The downsampled images are the outputs of the *LP_conv* layer, which applies a 2D LP filter, and the Gaussian_conv layer, which utilizes Gaussian filters with varying values of z and σ. These images are reduced to half the size of the original. The upsampled images are generated by the CNN layer using bilinear interpolation, increasing their size by a factor of two to match the original image dimensions.

From Fig. 3, we can make the following observations.

- ***LP_conv* based CNN block**: Effectively eliminate fine details, including noise, while retaining key structures.
- ***Gaussian_conv* based CNN block**:
  - *Filter size* (z): Smaller sizes are computationally efficient, reducing localized noise while preserving edges and fine details. Larger sizes provide broader smoothing, targeting noise over larger areas but with potential detail loss.
  - *Sigma* (σ): Lower values minimize smoothing, maintaining sharp edges. Higher values deliver more aggressive noise reduction, suitable for removing small noise but may blur finer details.

Both methods show effectiveness for removing small noise while preserving essential image features. The Gaussian filter, in particular, allows for targeted noise removal, offering a balance between computational efficiency and detail preservation.

3 Experiment

We integrate the designed CNN block into various DNN architectures, including ResNet18 [11], MobileNetV2 [12], and VGG16 [13]. These models are trained on the widely used German Traffic Sign Recognition Benchmark (GTSRB) dataset [14], which contains 43 classes of traffic signs, split into 39,209 training images and 12,630 test images. To evaluate the robustness of the trained models, we test them using the FGSM attack and a black-patch attack.

3.1 Adversarial attacks

3.1.1 FGSM attack

FGSM generates adversarial examples by slightly perturbing the input data in a way that maximizes the model's prediction error while keeping the perturbation imperceptible to humans. The attack works by exploiting the gradients of the loss function with respect to the input data. By taking a step in the direction of the gradient's sign, the attack aims to increase the loss and mislead the model into making incorrect predictions. The FGSM attack can be expressed mathematically as:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

The adversarial example x_{adv} is generated from the original input x using a perturbation factor ϵ to control the level of adversarial noise. The sign of the gradient of the loss J (calculated with respect to x) indicates the direction of modification. θ is the model parameter and y indicates the true label. FGSM is considered as a white-box attack as it requires prior knowledge of the neural network structure and parameters.

3.1.2 Black-box patch attack

In addition to the FGSM attack, we apply a black-patch attack that does not require knowledge of the DNN structure. This attack randomly places small black-box patches on images to simulate both real-world physical attacks (e.g., patches manually applied to traffic signs or cameras) and digital attacks (e.g., patches added to input data).

In the experiments, we used FGSM with ϵ values of 0.01, 0.05, 0.1, and 0.2. For the black-patch attack, we used black-box patches of size 3×3 pixels and varied the number of patches to 2, 4, 6, and 8 to simulate different levels of perturbations, as shown in Fig. 4.

3.2 Model performance under FGSM and black-box attacks

The proposed multi-resolution training method essentially integrates two key mechanisms: a low-pass filtering process and a downsampling-upsampling structure, both designed to mitigate adversarial perturbations by reducing high-frequency noise introduced by attacks. To assess their individual effectiveness, we also evaluate a variant of the proposed method, referred to as single-resolution training, which directly applies low-pass filtering within the CNN block while preserving the original image resolution. This comparison highlights the distinct contributions of filtering alone and the combined downsampling-upsampling process in enhancing adversarial robustness.

We developed multiple models by incorporating different configurations of our proposed multi-resolution and single-resolution filtering based CNN blocks into the base architectures of ResNet18, MobileNetV2, and VGG16. These models were evaluated under FGSM and black-box attacks with various levels for classification accuracy on the test dataset, as shown in Figs. 5, 6, and 7, respectively.

The prefixes and suffixes in the model names indicate specific modifications:

- Prefixes:
  - lpf_: Models integrated with the designed *LP_conv* based CNN block.
  - gs_: Models integrated with the designed *Gaussian_conv* based CNN block using a filter size of 3×3 and σ of 0.5
  - gm_: Models integrated with the designed *Gaussian_conv* based CNN block using a filter size of 5×5 and σ of 1.5.
  - gl_: Models integrated with the designed *Gaussian_conv* based CNN block using a filter size of 7×7 and σ of 2.5.
- Suffixes:
  - _c3: The CNN block outputs a processed 3-channel image, which is then fed into the main DNNs for processing.
  - _c6: The CNN block concatenates the processed 3-channel image with the original image to create a 6-channel input for the main DNNs (the DNNs are adjusted to accept a 6-channel input for compatibility).

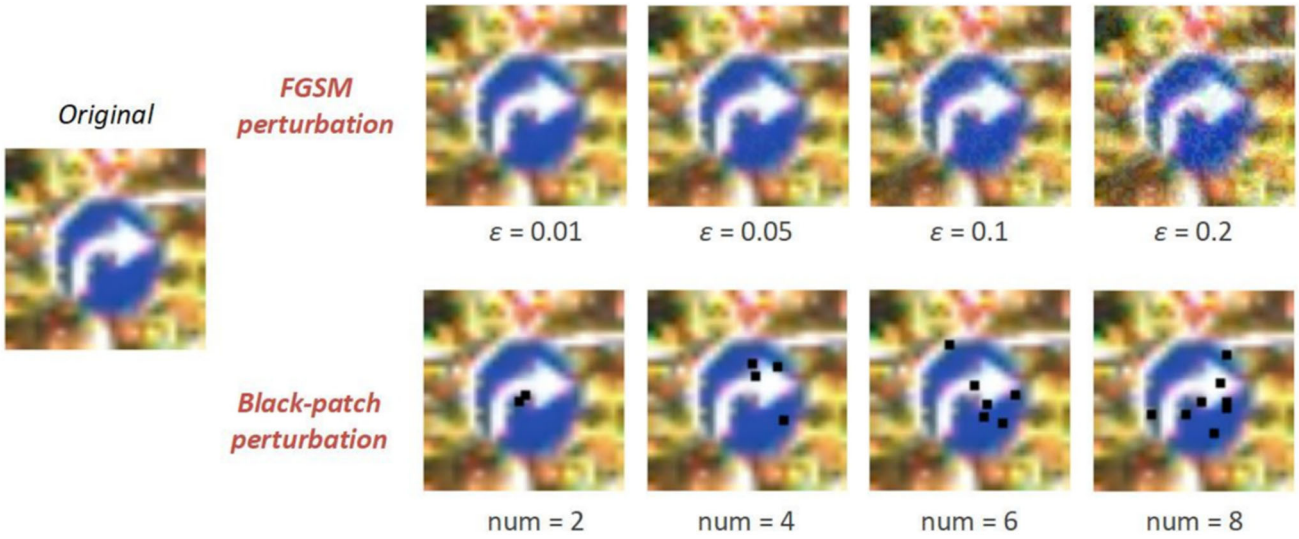


Fig. 4 Examples of FGSM and black-patch perturbations

Figures 5, 6 and 7 illustrate the impact of FGSM and black-box attacks on ResNet18, MobileNetV2, and VGG16, with comparisons between standard models and their enhanced versions incorporating LPF and Gaussian-based (gs_, gm_, gl_) filtering CNN blocks. These enhancements effectively improve model robustness by suppressing high-frequency adversarial noise while preserving essential structural information.

3.2.1 ResNet18-based models (Fig. 5)

In Fig. 5a, as FGSM attack level increases, all models shows accuracy declines, with accuracy dropping from approximately 0.7−0.8 at $\epsilon = 0.01$ to around 0.2 at $\epsilon = 0.2$. However, models enhanced by multi-resolution CNN block (particularly gl_resnet18_c3 model) consistently outperform their base models by 6–12%. The effectiveness of Gaussian-based filtering in ResNet18 highlights that adversarial perturbations mainly rely on high-frequency textures, and applying a 7×7 Gaussian filter (gl_) removes this noise while retaining global structures, improving robustness against adversarial gradients.

In Fig. 5b, as the number of adversarial boxes increases from 2 to 8, all models experience a decline in accuracy, dropping from approximately 0.85 to around 0.60. The proposed filtering techniques provide a consistent improvement of 5–8%, indicating that enhanced models (particularly gl_resnet18_c3 model) are more robust against black-box perturbations by eliminating high-frequency distortions without over-blurring key image details.

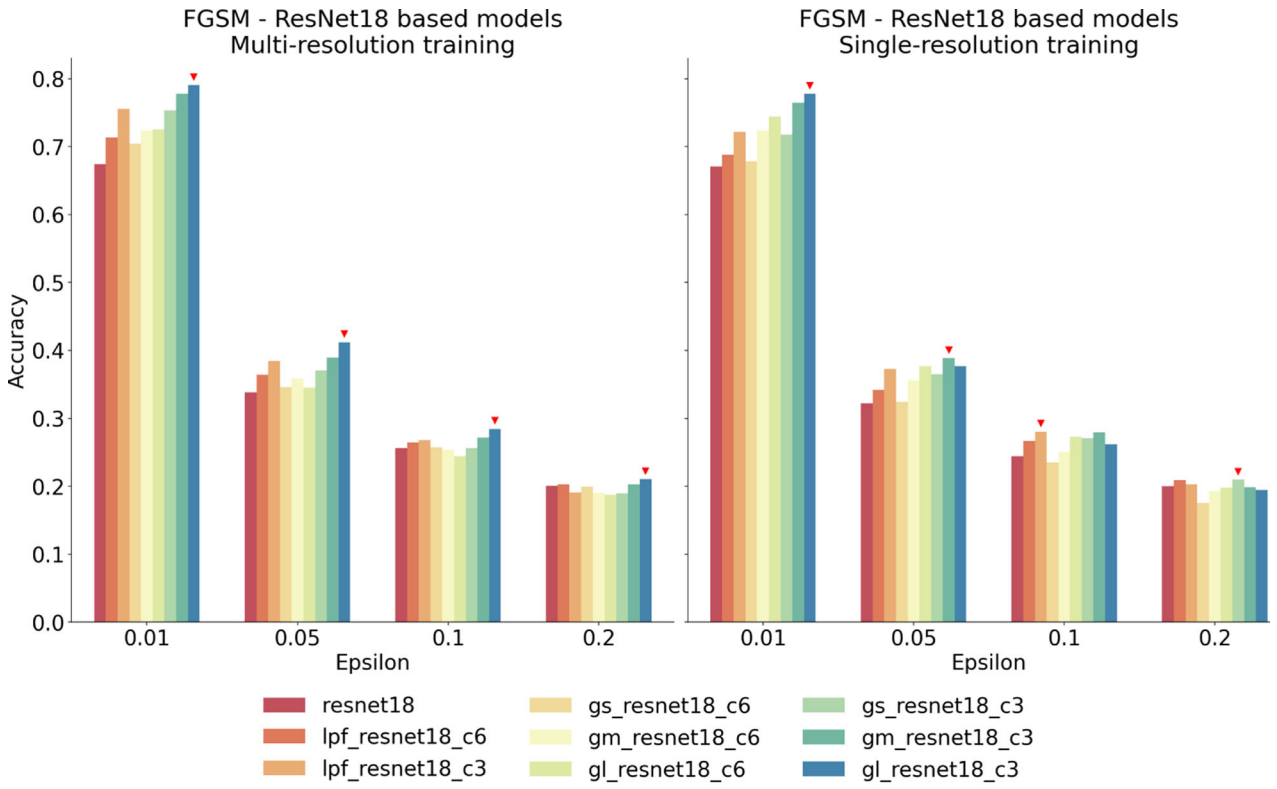
3.2.2 MobileNetV2-based models (Fig. 6)

For FGSM attacks in Fig. 6a, MobileNetV2 exhibits a steeper accuracy drop than ResNet18, decreasing from 0.78 to 0.20 at $\epsilon = 0.2$. However, multi-resolution training improves accuracy by almost 18% when $\epsilon = 0.01$ and maintains around 5-10% accuracy improvement when ϵ increases. This demonstrates that although MobileNetV2 is highly susceptible to high-frequency adversarial noise, but filtering can effectively mitigate this issue. The results further emphasize that gl_ enhanced models outperform others, suggesting that deeper smoothing reduces sensitivity of MobileNetV2 to small-scale perturbations.

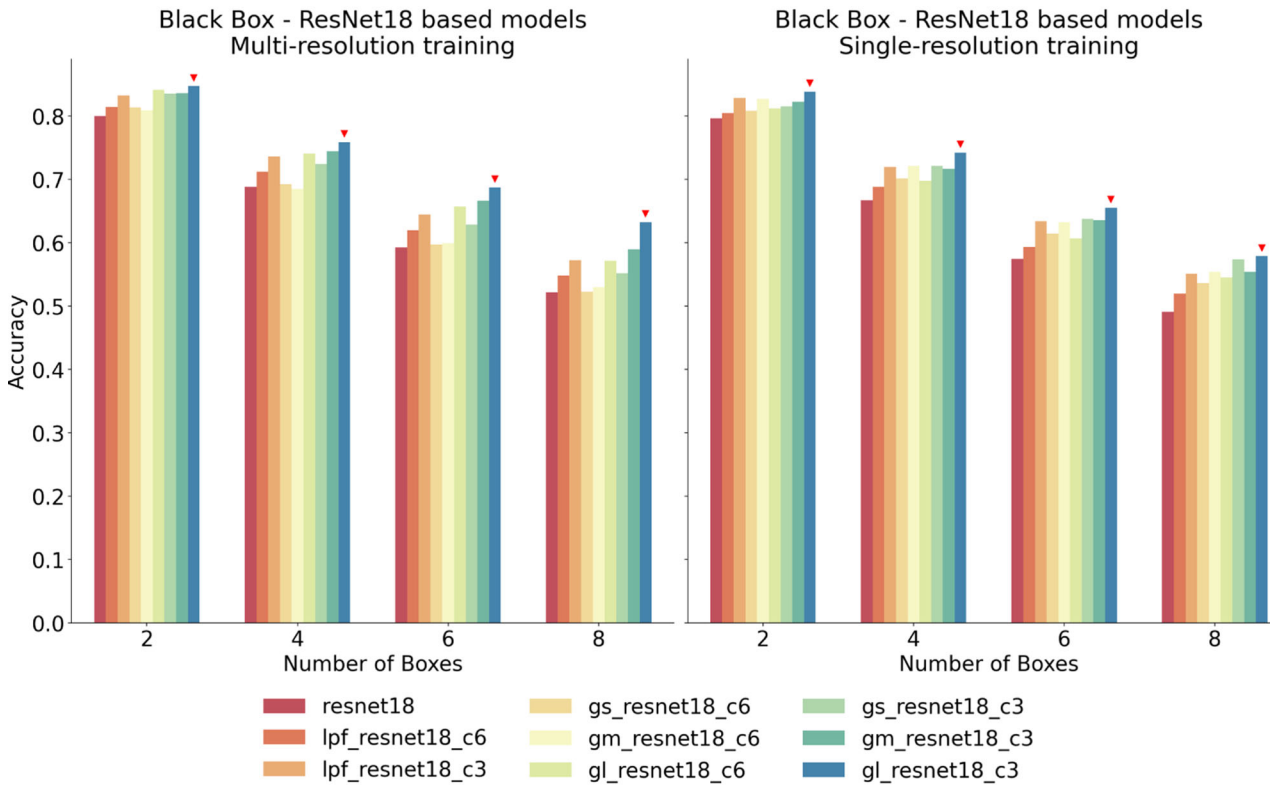
Black-box attacks in Fig. 6b show that accuracy drops from approximately 0.85–0.65 as the number of adversarial patches increases. Interestingly, MobileNetV2 benefits less from filtering techniques under black-box attacks (around 2% improvement) compared to FGSM (5–20%). This is due to its depthwise separable convolutions, which process spatial and channel-wise information separately. This architecture inherently providing robustness against structured, lower-frequency perturbations introduced by black-box attacks.

3.2.3 VGG16-based models (Fig. 7)

Figure 7a shows that VGG16 is the most resilient model against FGSM due to its deeper convolutional structure, maintaining accuracy above 30% even at $\epsilon = 0.2$, outperforming ResNet18 and MobileNetV2. The strongest enhancement occurs in gl_vgg16_c3 based models, where multi-resolution training improves accuracy by 10% compared to standard model when $\epsilon = 0.01$. This is followed by the lpf_vgg16_c6

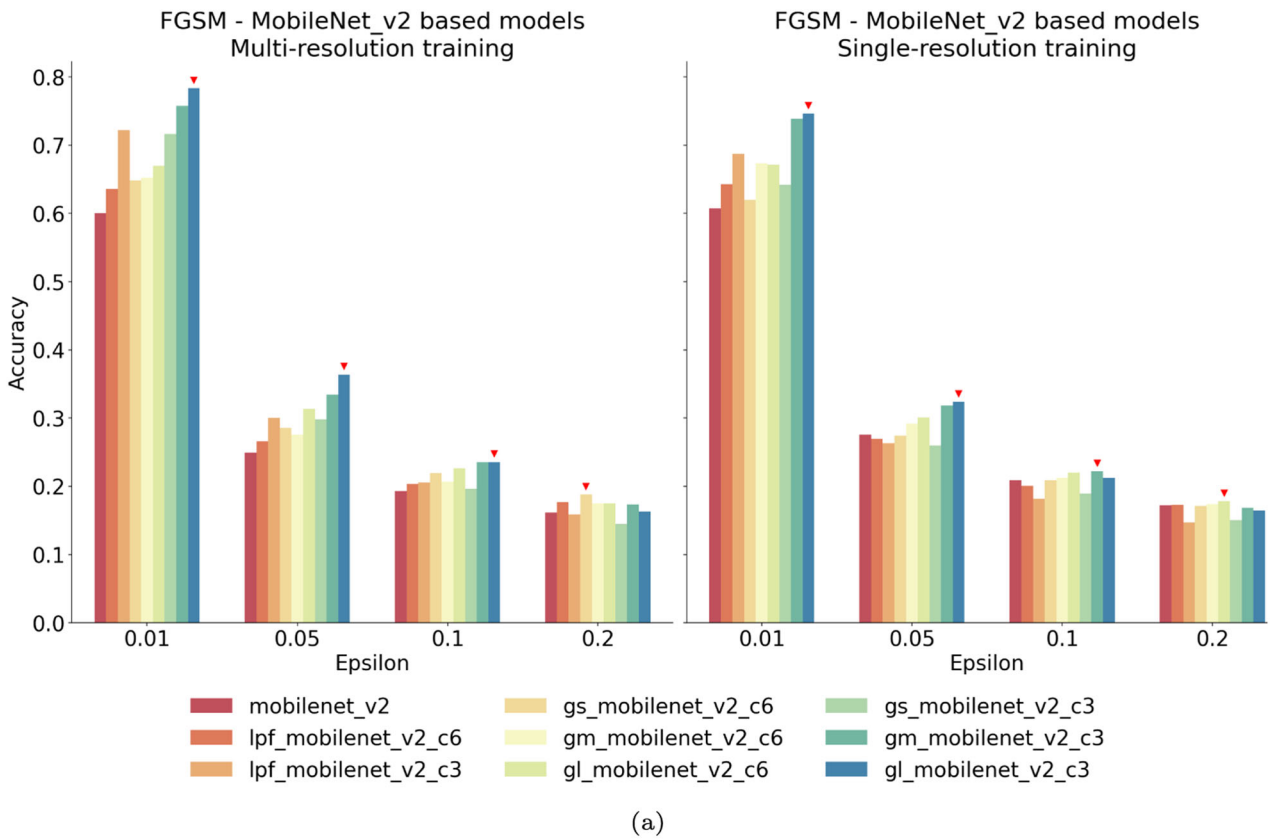


(a)

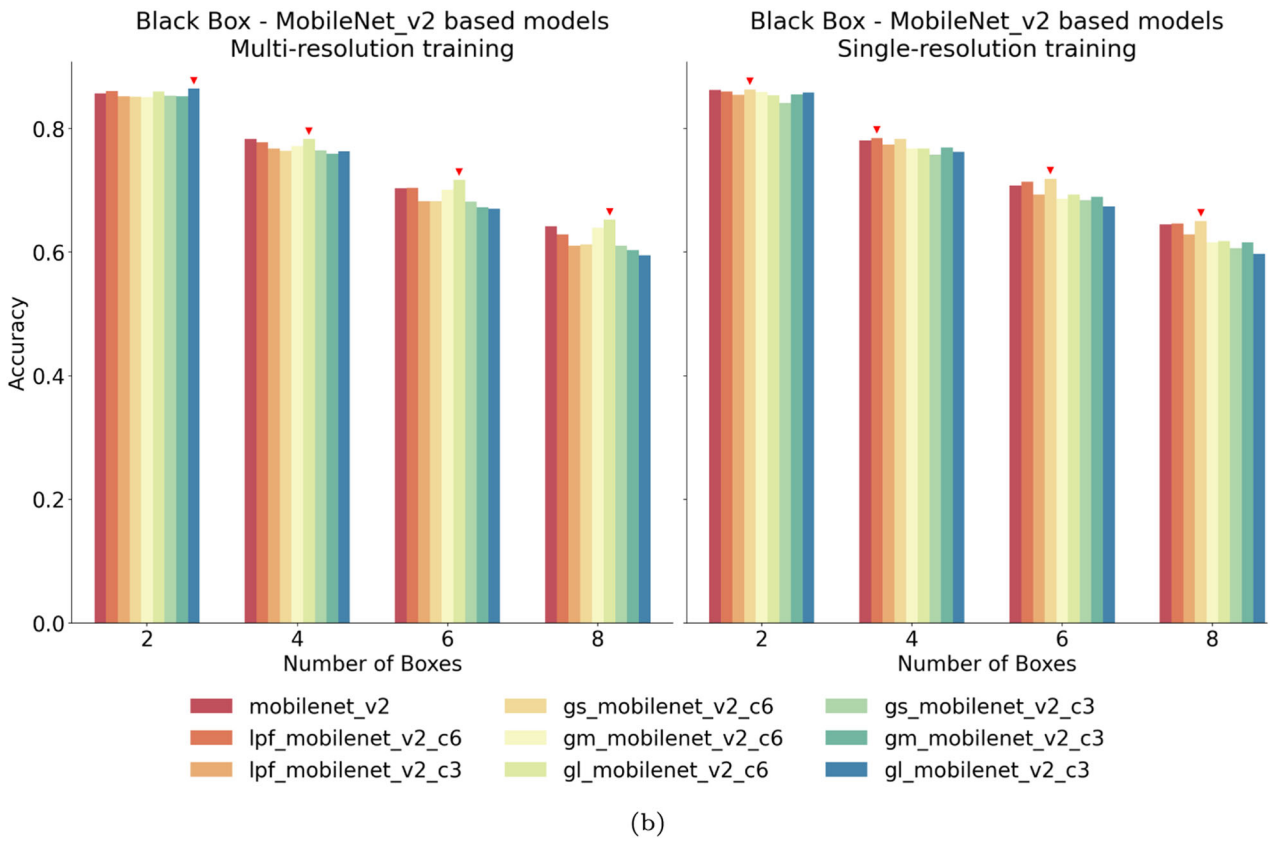


(b)

Fig. 5 Model accuracy under attack on ResNet18-based models. **a** FGSM attack. **b** black-box attack

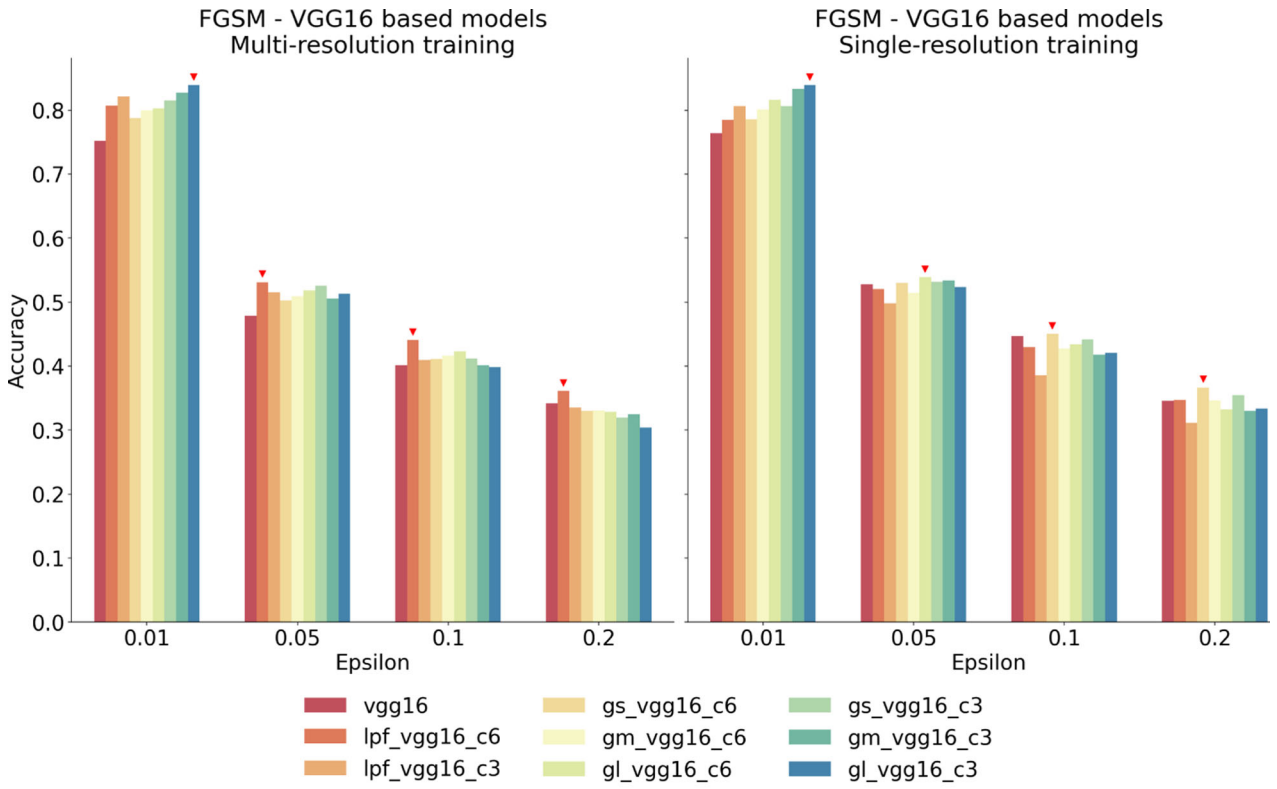


(a)

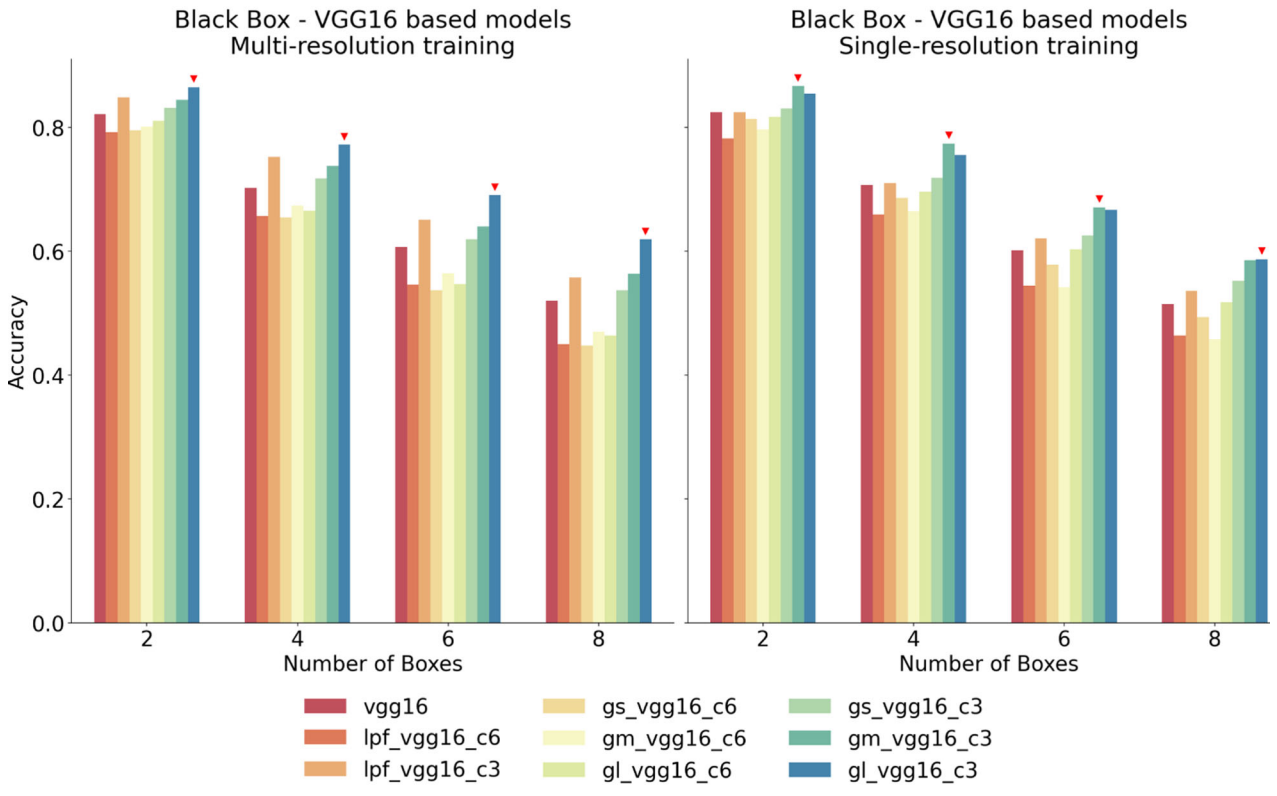


(b)

Fig. 6 Model accuracy under attack on MobileNetV2-based models. **a** FGSM attack. **b** black-box attack



(a)



(b)

Fig. 7 Model accuracy under attack on VGG16-based models. **a** FGSM attack. **b** Black-Box attack.

models which provide approximately 4-7% improvement as ϵ increases.

Figure 7b highlights that VGG16 also benefits from filtering techniques against black-box attacks, though not all methods show consistent improvements. Filtering models with the _c6 configuration, including lpf_vgg16_c6 and gs/m/l_vgg16_c6, show no enhancement. However, _c3 models, particularly gl_vgg16_c3, remains prominent, achieving an accuracy improvement of approximately 4–10%. This reinforces that filtering-based enhancements strengthen VGG16 robustness against both gradient-based (FGSM) and transfer-based (black-box) attacks.

3.3 Key takeaways from results

Multi-resolution training is consistently more effective than single-resolution training, particularly at higher attack strengths (ϵ = 0.2 or 8 adversarial black-box patches). This confirms that multi-scale feature extraction prevents adversarial attacks from exploiting a single frequency band, improving overall robustness. gl_ filtered models (7×7, σ = 2.5) generally provide the most substantial improvement, suggesting that aggressive Gaussian smoothing removes adversarial distortions more effectively than smaller-scale smoothing.

VGG16 demonstrates strong baseline robustness, benefiting from deep feature extraction and redundancy, but it experiences a sharper decline under stronger black-box attacks (Fig. 7b. ResNet18 is highly vulnerable to FGSM attacks (Fig. 5a) due to residual connections that propagate adversarial gradients, yet it gains the most from proposed filtering CNN block, particularly with gl_ configuration. MobileNetV2 also shows vulnerability to FGSM (Fig. 6a) but is more stable under black-box attacks (Fig. 6b), likely due to its depthwise separable convolutions reducing sensitivity to structured perturbations.

FGSM attacks cause more severe accuracy degradation than black-box attacks, due to FGSM ability to manipulate model gradients to generate optimized adversarial perturbations. In contrast, black-box attacks introduce more generic, less optimized perturbations, leading to less accuracy degradation.

4 Key insights on proposed work

Effectiveness of enhancements: The enhanced models incorporating LP and Gaussian filtering-based CNN blocks consistently outperform their respective baselines (ResNet18, MobileNetV2, and VGG16) across both FGSM and black-box attacks. This highlights the effectiveness of the proposed filtering mechanisms in improving model robustness. The performance gains observed across Figs. 5, 6, and 7 confirm that enhancing convolutional networks with frequency-based filtering techniques provides a strong defense against adversarial perturbations, particularly in models that struggle with gradient-based adversarial attacks.

Impact of c6 versus c3 configurations: Another key observation from the study is that the c3 configuration generally outperforms the c6 configuration. This indicates that directly processing the 3-channel input through the designed CNN block and feeding it into the main deep network (c3) is more effective than concatenating it with the original image (c6). The superior performance of c3 suggests that standalone processed features provide sufficient robustness without requiring additional raw feature information. This further confirms that filtering-based feature transformation is effective in mitigating adversarial distortions without losing critical image information.

Performance against adversarial attacks: As adversarial perturbations exacerbate, whether through increasing ϵ in FGSM or adding more patches in black-box attacks, the robustness of enhanced models typically decreases. However, models equipped with gl_ blocks and c3 configurations still outperform baseline models, maintaining higher accuracy under attack conditions. Nonetheless, results suggest that this combination shows a quicker decline in performance as perturbations grow, highlighting that while filtering improves initial robustness, its effectiveness diminishes against extreme adversarial perturbations.

Multi-resolution versus single-resolution training: Figure 8 presents a direct comparison of accuracy improvements between multi-resolution and single-resolution training, based on the average accuracy gains across different attack levels. Multi-resolution training consistently provides higher accuracy gains across all models and attack types. In ResNet18, multi-resolution training improves FGSM accuracy by 5.7% and black-box accuracy by 8.1%, while single-resolution training achieves 5.5% and 7.1%, respectively. MobileNetV2 exhibits the greatest improvement under FGSM attacks (9.2% in multi-resolution settings), confirming that filtering effectively enhances model robustness against high-frequency adversarial perturbations. However, its black-box improvement is minimal (about 1%) due to its inherent ability to counter black-box attacks. VGG16 benefits from multi-resolution training technique under both attack types, achieving improvements in FGSM (4.9%) and black-box (7.4%) attack scenarios. Overall, multi-resolution training consistently outperforms single-resolution training, demonstrating that extracting features at multiple resolution scales enhances adversarial robustness.

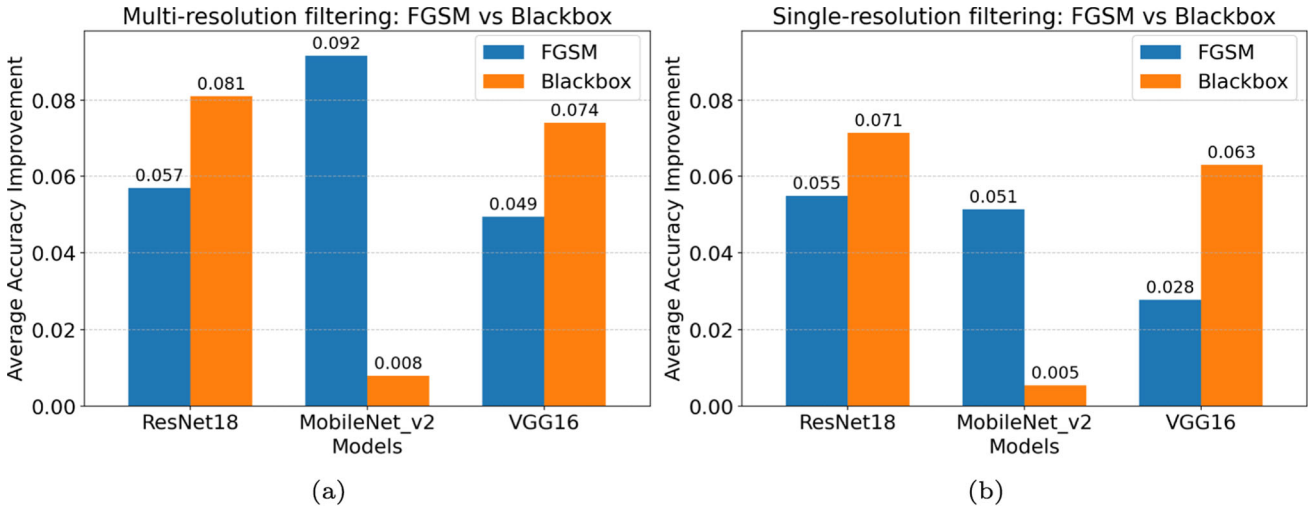


Fig. 8 Average accuracy improvement over various attack levels: **a** Multi-resolution training. **b** Single-resolution training

5 Conclusion

DNNs are essential for applications like traffic sign recognition, but their vulnerability to adversarial attacks poses significant deployment challenges. This paper introduced a multi-resolution training approach that enhances model robustness by integrating custom LP and Gaussian filtering-based CNN blocks, offering an effective and scalable method for adversarial training. This approach can be used together with the traditional adversarial training approaches to achieve enhanced robustness against adversarial attacks.

Key findings:Multi-resolution training effectively enhances model robustness against adversarial attacks, with `gl_` filtering ($7\times7, \sigma = 2.5$) offering the most robustness enhancement. When combined with the c3 configuration, it achieves the highest adversarial accuracy improvement, improving performance by up to 18% (e.g., MobileNetV2 under mild FGSM attacks).

Multi-resolution training benefits VGG16, ResNet18, and MobileNetV2, each to varying degrees. VGG16 and ResNet18 exhibit balanced improvements, enhancing robustness against both FGSM and black-box attacks, with an average accuracy gain of 6-8%. MobileNetV2 benefits most under FGSM, achieving an average 9% accuracy improvement, but gains less from black-box attacks due to its depthwise separable convolutions providing inherent robustness to structured perturbations.

These findings emphasize the broad applicability and adaptability of multi-resolution training in enhancing adversarial robustness across different DNN architectures.

Comparison with adversarial training:Traditional adversarial training enhances model robustness by augmenting training data with adversarial examples but struggles with generalization to unseen attacks, and requires significant computational efforts. This makes it resource-intensive and impractical for large-scale applications.

In contrast, our multi-resolution training approach does not rely on adversarial data augmentation. Instead, it strengthens model robustness by suppressing high-frequency adversarial noise at the feature extraction level, making it computationally efficient and scalable across different architectures. Unlike adversarial training, which tends to overfit to specific attack patterns, multi-resolution training generalizes across different perturbation types, improving robustness with lower training costs.

Limitations and future directions: While multi-resolution training proves effective, it still has limitations that need further exploration.

Diminished effectiveness under extreme perturbations: Filtering mitigates adversarial noise, but its performance declines as attack intensity increases. In the future, we will explore adaptive filtering techniques that dynamically adjust filter strength based on input perturbation levels.

Defense against stronger adversarial attacks: Our future research will explore defenses against more complex adversarial attacks, including adaptive attacks that attempt to bypass filtering mechanisms.

The results highlight multi-resolution training as a viable alternative or complement to adversarial training, enhancing DNN robustness without requiring adversarial data augmentation. By removing high-frequency adversarial noise while preserving essential image features, this method effectively improves model robustness against both FGSM and black-box attacks.

These findings pave the way for safer and more reliable deep learning applications, particularly in real-world safety-critical systems, such as autonomous driving, medical diagnostics, and security-based AI models. Our future

research will focus on further optimizing multi-resolution training techniques and expanding its applications to broader AI domains.

References

1. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. International Conference on Learning Representations (ICLR) (2014)
2. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR) (2015)
3. Pavlitska, S., Lambing, N., Zöllner, J.M.: Adversarial attacks on traffic sign recognition: A survey. 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (2023)
4. Eykholt, K., et al.: Robust physical-world attacks on deep learning visual classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
5. Zolfi, A., Kravchik, M., Elovici, Y., Shabtai, A.: The translucent patch: A physical and universal attack on object detectors. Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR) (2018)
7. Papernot, N., McDaniel, P.D., Goodfellow, I.J., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. Asia Conference on Computer and Communications Security (AsiaCCS) (2017)
8. Papernot, N., McDaniel, P.D., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. IEEE European Symposium on Security and Privacy (EuroS&P) (2016)
9. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy (SP) (2017)
10. Zhang, Y., Cui, J., Liu, M.: Research on adversarial patch attack defense method for traffic sign detection. In: Cyber Security. CNCERT 2022. Communications in Computer and Information Science, pp. 199–210. Springer, Singapore (2022). doi: https://doi.org/10.1007/978-981-19-8285-9_15
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2015). arXiv preprint arXiv:1512.03385
12. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks (2019). arXiv preprint arXiv:1801.04381
13. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2015). arXiv preprint arXiv:1409.1556
14. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: A multi-class classification competition. In: IEEE International Joint Conference on Neural Networks, pp. 1453–1460 (2011)