Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34237

Feature-Dependent Generative Adversarial Network for Robustness Testing of Image Classification Models

Wenqi He

School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China
3014386191@qq.com

Abstract: Image classification models with high benchmark accuracy may still be vulnerable to small structured perturbations, making robustness testing essential before deployment. This paper proposes a feature-dependent generative adversarial network for generating robustness testing samples. The method extracts deep visual features from the input image, fuses them with a random latent vector, and decodes the fused representation into a residual perturbation. The perturbation is added to the original image to produce a visually similar test sample. The training objective combines image reconstruction loss, classification loss, and feature loss to preserve image semantics while reducing classifier confidence and disrupting internal representations. Experiments on ImageNet with a ResNet50 classifier show that the proposed method reduces accuracy from 76% to 66% under an image loss of 0.003. The ablation study confirms that the feature extraction module improves attack effectiveness, demonstrating the feasibility of feature-dependent perturbation generation for robustness testing.

Keywords: adversarial examples, generative adversarial network, robustness testing, feature loss, image classification

1. Introduction

Deep convolutional neural networks have become a standard route for visual recognition, and their performance on datasets such as ImageNet has supported applications in object recognition, content moderation, medical imaging, and autonomous systems. Clean-set accuracy, however, does not fully describe reliability in open environments. A classifier can retain high validation accuracy while producing unstable predictions under noise, compression artifacts, sensor variation, or adversarial perturbation. The existence of adversarial examples shows a gap between human semantic perception and learned decision boundaries [1].

Robustness testing focuses on this gap. For image classification, a useful testing sample should remain visually close to the original image while causing measurable degradation in the target model. Gradient-based attacks such as FGSM construct such samples by moving the input in the direction that increases classification loss [2]. These methods are effective, but they usually require access to the architecture, parameters, and gradients of the target classifier. This dependence limits their use when the evaluator needs reusable samples rather than repeated white-box optimization.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34237

Generator-based methods provide a complementary route. Once a generator has learned how to synthesize perturbations, new testing samples can be produced through a forward pass. This paper proposes FD-GAN, a feature-dependent generator for robustness testing. The model extracts deep features from an input image, concatenates them with a random latent vector, and decodes the fused representation into an image-sized residual perturbation. The training objective combines image similarity, misclassification pressure, and feature disruption.

The contribution of this paper is a compact testing framework that integrates residual generation, feature conditioning, and representation-level loss. The method is not designed to replace the strongest white-box attack; rather, it provides a reusable generator that can produce visually constrained robustness testing samples after training. Experiments on ImageNet with a ResNet50 classifier and an ablation study verify the role of feature-dependent conditioning.

2. Related works

Adversarial examples show that deep neural networks can be sensitive to small perturbations. Szegedy et al. demonstrated that such perturbations can change predictions and sometimes transfer across models [1]. Goodfellow et al. proposed FGSM and explained this vulnerability through the high-dimensional linear behavior of neural networks [2]. Since then, robustness evaluation has moved beyond clean accuracy toward model behavior under constrained input changes.

Gradient-based attacks, including FGSM, projected-gradient methods, robust optimization, and momentum attacks, remain strong tools for robustness evaluation [2-4]. However, their sample generation is usually coupled with the target model or a substitute model. Generator-based attacks, such as AdvGAN, reduce per-sample optimization by learning a perturbation synthesis network [5]. This makes generator-based methods more suitable for reusable robustness testing after training.

GANs learn generation through an adversarial game [6], while conditional GANs improve controllability by injecting auxiliary information [7]. In adversarial testing, the condition can be a class label, image representation, or target feature. FD-GAN follows this conditional generation idea but uses the deep feature vector of the input image as the condition, making the generated perturbation sample-dependent.

Feature-space attacks further motivate the proposed feature loss. Prior studies show that perturbing intermediate representations can improve transferability and affect model behavior beyond the final logits [8, 9]. Recent feature-aware, black-box feature, multi-model transfer, and attentional semantic attacks also confirm that representation-level information is important for transferable adversarial generation [10-14]. Therefore, FD-GAN uses feature guidance to generate repeatable robustness testing samples rather than relying only on output-level loss or per-image gradient optimization.

3. Proposed methodology

3.1. Problem formulation

Let $x \in \mathbb{R}^{H\times W\times C}$ denote an input image and y denote its ground-truth class. A classifier F_θ predicts the class distribution $F_\theta(x)$, and the predicted label is obtained by

$$\hat{y} = \mathrm{argmax}_k F_\theta(x)_k. \tag{1}$$

DOI: 10.54254/2755-2721/2026.34237

The goal of robustness testing sample generation is to construct x' such that x' is visually close to x, while $F_\theta(x')$ is more likely to produce an incorrect prediction. In this paper, the generated sample is defined as

$$x' = x + \delta, \quad \delta = G_\phi(h(x), z), \tag{2}$$

where G_ϕ is the perturbation generator, $h(x)$ is the feature vector extracted from the input image, and z is a random latent vector. The perturbation δ is constrained through the image loss and the generator architecture rather than through a single manually designed pixel-level rule.

The perturbation is treated as a residual map rather than a new image. This design is appropriate for robustness testing because the original semantic content is preserved by construction. The generator learns where and how to add small perturbations that are more likely to disturb the target classifier.

3.2. Network architecture

The overall architecture of FD-GAN is shown in Figure 1. The input image is first processed by a ResNet50-based feature extraction module. The extracted feature representation is concatenated with a random vector and then projected into a low-resolution tensor. A sequence of up-convolutional layers restores the tensor to the original image resolution and generates the perturbation map. The perturbation is added to the original image to form the robustness testing sample.

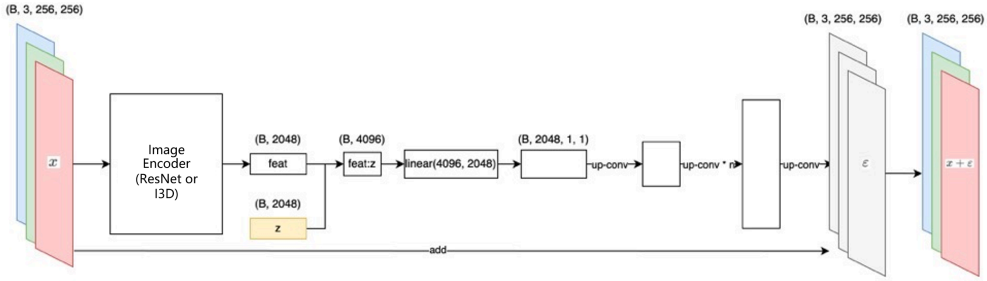


Figure 1. Architecture of the proposed feature-dependent perturbation generator

Different from a conventional GAN generator that maps only random noise to an image, FD-GAN conditions perturbation generation on the feature representation of the specific input image. In this study, the discriminator adopts a ResNet50 backbone with a fully connected classification head, and the generator uses the ResNet50 feature output as $h(x)$. The feature vector $h(x)$ is concatenated with a random vector z, which introduces stochasticity and avoids a deterministic one-to-one mapping from image to perturbation. The fused representation is decoded into the perturbation map δ, and the generated testing sample is obtained by adding δ to the original image. This residual formulation preserves the semantic content of the input while encouraging the generator to learn minimal but effective visual modifications.

The main innovation of FD-GAN lies in the joint use of residual generation, feature conditioning, and representation disruption. Residual generation keeps the generated sample close to the original image, feature conditioning injects classifier-related semantic information into the generator, and representation disruption extends the attack target from the final logits to the intermediate feature

DOI: 10.54254/2755-2721/2026.34237

space [12, 13]. To constrain perturbation magnitude and improve training stability, the raw decoder output $\tilde{\delta}$ is bounded as:

$$\delta = \epsilon \cdot tanh\left(\tilde{\delta}\right), \tag{3}$$

where ϵ controls the maximum perturbation amplitude. The generated image is then clipped to the valid pixel interval:

$$x' = \text{clip}\left(x + \delta, 0,1\right). \tag{4}$$

The generator can also be extended with a feature attention gate. Given $h(x)$, an attention weight a is produced by

$$a = \sigma(W_a h(x) + b_a), \tag{5}$$

and the fused representation is written as

$$u = [a \odot h(x); z]. \tag{6}$$

This attention gate allows the generator to emphasize feature channels that are more sensitive to classification behavior. In the current implementation, direct feature concatenation is used for simplicity, while the attention-gated form provides a natural extension for future experiments.

3.3. Loss function and optimization

The training objective contains three main terms: image loss, classification loss, and feature loss. Image loss constrains the distance between x and $x\prime$. Classification loss encourages the generated sample to increase the target classifier error. Feature loss enlarges the discrepancy between $h(x)$ and $h(x')$, forcing the perturbation to affect internal representations rather than only the output layer.

The total objective is:

$$\mathcal{L} = \lambda_{img}\mathcal{L}_{img} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{feat}\mathcal{L}_{feat}. \tag{7}$$

The image term preserves visual semantics, the classification term drives misclassification, and the feature term disrupts the representation used by the classifier. The weights control the trade-off between visual quality and attack effect. In implementation, a total variation regularizer can also be added to reduce isolated high-frequency artifacts.

During training, each mini-batch image is encoded into $h(x)$, fused with z, decoded into δ, and added to x. The fixed classifier evaluates x' , and only the generator is updated. During inference, FD-GAN performs feature extraction and perturbation generation with one forward pass. This separation between training and inference makes the method suitable for repeated batch robustness testing.

The weights of the loss function should therefore be selected according to the testing goal. If visual fidelity is the priority, λ_{img} should be larger and the generated samples will be conservative. If the goal is to expose model vulnerability, λ_{cls} and λ_{feat} can be increased, but the image loss must

DOI: 10.54254/2755-2721/2026.34237

still be monitored. In a formal evaluation, these weig13hts should be reported together with the resulting perturbation budget so that different methods can be compared under similar visual constraints.

4. Experiments and analysis

4.1. Dataset and metrics

The experiment uses ImageNet, a large-scale classification benchmark with 1000 categories [15]. The target model is ResNet50 with a fully connected classification head [16]. The reported clean validation accuracy is 76%. Two primary metrics are used: classification accuracy and image loss. Lower accuracy indicates stronger robustness testing samples, while lower image loss indicates smaller visual distortion.

Although the current experiment uses accuracy and image loss, the framework can be extended with attack success rate, confidence drop, SSIM, PSNR, or LPIPS. These metrics would provide a more complete view of whether generated samples satisfy both machine-level attack effectiveness and human-level visual similarity. They are not included in the current source experiment, but they are compatible with the proposed evaluation framework.

4.2. Main results

Table 1 compares clean input, FGSM, FD-GAN, and the ablated FD-GAN without the feature module. FGSM is stronger when white-box gradients are available, but FD-GAN provides a generator-based alternative that produces samples through forward propagation after training. Under the same image loss of 0.003, FD-GAN reduces accuracy from 76% to 66%, which shows that the generated residual perturbation can expose classifier sensitivity while keeping the image visually constrained.

The comparison clarifies the intended role of FD-GAN. It trades part of the attack strength for reusable generation. In practical model evaluation, this balance can help identify models that appear accurate on clean validation data but become unstable under small feature-dependent residuals.

Table 1. Classification accuracy of ResNet50 on clean and generated samples

Method	Image loss	Accuracy
Clean input	0	76.0%
FGSM	0.050	32.8%
FGSM	0.003	45.1%
FD-GAN (ours)	0.003	66.0%
FD-GAN without feature module	0.004	71.0%

DOI: 10.54254/2755-2721/2026.34237

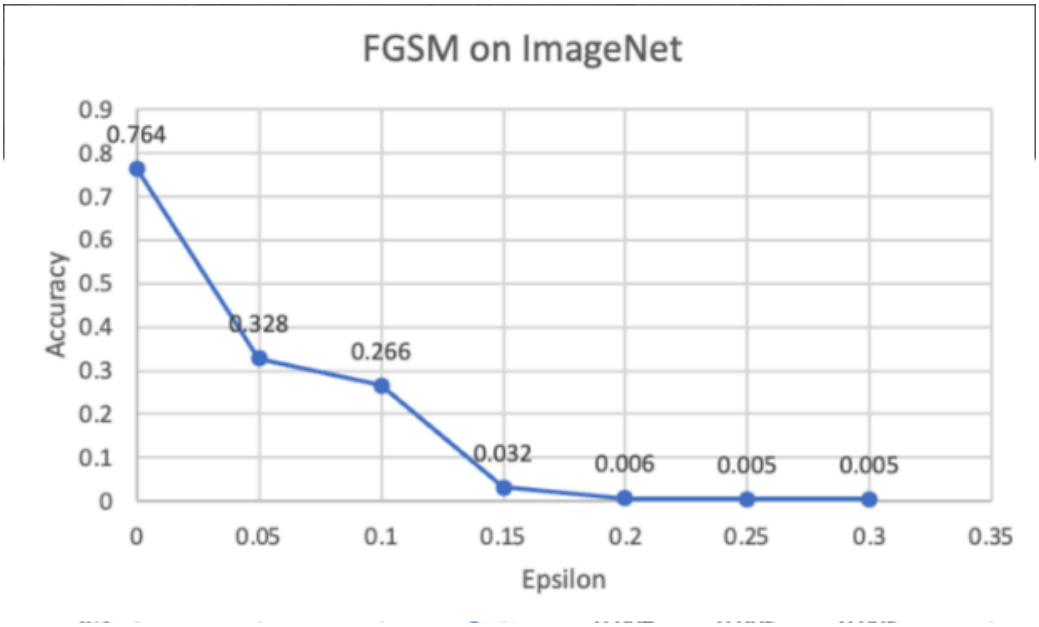


Figure 2. Accuracy of the ImageNet classifier under FGSM perturbations

The result should be interpreted together with the method objective. FGSM is expected to be stronger because it directly uses gradient information from the target model. FD-GAN instead emphasizes forward generation after training. Therefore, the observed 10 percentage-point accuracy drop under a small image loss indicates that the generator has learned a meaningful perturbation distribution, even though it does not maximize attack strength in the same way as a white-box optimizer.

4.3. Ablation and visual analysis

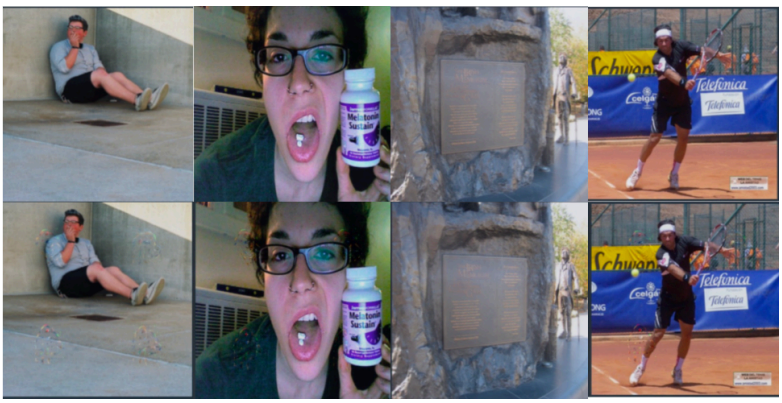


Figure 3. Visual comparison between an original image and the generated robustness testing sample

The ablation removes the feature extraction module from the generator. Accuracy only decreases to 71%, and image loss increases to 0.004. Compared with the full FD-GAN result, the ablation confirms that feature conditioning provides useful image-specific guidance for perturbation generation.

Figure 3 compares original images (first row) and generated samples (second row). The generated images preserve the main objects and scene structures, while local perturbations remain limited to

DOI: 10.54254/2755-2721/2026.34237

specific regions. This behavior is consistent with the residual design: FD-GAN does not generate a new image from scratch but adds a learned perturbation to the original input. The visual evidence matches the reported image loss and supports the semantic-preservation requirement of robustness testing.

4.4. Discussion

The experimental results show a trade-off between attack strength, model access, and visual constraint. FGSM is stronger in reducing target accuracy because it directly follows the gradient of the target classifier. FD-GAN is weaker under the same image-loss value, but its generation process is more suitable for efficient repeated testing after training. The ablation result also indicates that feature dependence is not decorative; it provides useful guidance for perturbation generation.

The current experiment mainly validates feasibility on a ResNet50 target classifier. Since the method is motivated by feature-level disruption and transferability, future experiments should evaluate generated samples on different architectures, such as VGG, DenseNet, Inception, and vision transformers. Additional perceptual metrics and generation-time measurements are also needed before the method can be positioned as a complete robustness testing benchmark.

5. Conclusion

We presents FD-GAN for robustness testing of image classification models. It extracts deep features, fuses them with latent noise, and generates residual perturbations optimized by image, classification, and feature losses. Experiments on ImageNet with ResNet50 show that FD-GAN reduces accuracy from 76% to 66% at an image loss of 0.003, while the ablation result confirms the contribution of the feature module. Overall, FD-GAN provides a reproducible feature-conditioned robustness testing framework and can be extended with ensemble feature extraction, attention-guided perturbation, and multi-model transfer evaluation.

References

[1] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R 2014 Intriguing properties of neural networks *Int. Conf. on Learning Representations*
[2] Goodfellow I J, Shlens J and Szegedy C 2015 Explaining and harnessing adversarial examples *Int. Conf. on Learning Representations*
[3] Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A 2018 Towards deep learning models resistant to adversarial attacks *Int. Conf. on Learning Representations*
[4] Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X and Li J 2018 Boosting adversarial attacks with momentum *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 9185-93
[5] Xiao C, Li B, Zhu J Y, He W, Liu M and Song D 2018 Generating adversarial examples with adversarial networks *Proc. 27th Int. Joint Conf. on Artificial Intelligence* pp 3905-11
[6] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Adv. Neural Inf. Process. Syst. 27* 2672-80
[7] Mirza M and Osindero S 2014 Conditional generative adversarial nets Preprint arXiv: 1411.1784
[8] Inkawhich N, Wen W, Li H H and Chen Y 2019 Feature space perturbations yield more transferable adversarial examples *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 7066-74
[9] Ganeshan A, Vivek B S and Babu R V 2019 FDA: Feature disruptive attack *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 8069-79
[10] Cheng S, Li P, Han K and Xu H 2024 Feature-aware transferable adversarial attacks against image classification *Appl. Soft Comput. 161* 111729
[11] Wang M, Wang J, Ma B and Luo X 2024 Improving the transferability of adversarial examples through black-box feature attacks *Neurocomputing 595* 127863

DOI: 10.54254/2755-2721/2026.34237

[12] Wang Y, Liao X, Cui W and Yang Y 2025 Defending against and generating adversarial examples together with generative adversarial networks *Sci. Rep. 15* 12994
[13] Zhu P, Jin X, Jiang Q, Gao X, Wang P, Yao S and Zhou W 2025 Transferable adversarial attacks for multi-model systems coupling image fusion with classification models *Cybersecurity 8* 23
[14] Wang P and Liu J 2026 Attentional semantic attack for enhancing adversarial samples transferability *Sci. Rep. 16* 10957
[15] Russakovsky O et al 2015 ImageNet large scale visual recognition challenge *Int. J. Comput. Vis. 115* 211-52
[16] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770-8