Applied Intelligence (2024) 54:10608–10626
https://doi.org/10.1007/s10489-024-05728-z

# Improving the transferability of adversarial attacks via self-ensemble

Shuyan Cheng[1] · Peng Li[1] · Jianguo Liu[1] · He Xu[1] · Yudong Yao[2] · *Fellow, IEEE*

Accepted: 4 August 2024 / Published online: 22 August 2024

## Abstract

Deep neural networks have been used extensively for diverse visual tasks, including object detection, face recognition, and image classification. However, they face several security threats, such as adversarial attacks. To improve the resistance of neural networks to adversarial attacks, researchers have investigated the security issues of models from the perspectives of both attacks and defenses. Recently, the transferability of adversarial attacks has received extensive attention, which promotes the application of adversarial attacks in practical scenarios. However, existing transferable attacks tend to trap into a poor local optimum and significantly degrade the transferability because the production of adversarial samples lacks randomness. Therefore, we propose a self-ensemble-based feature-level adversarial attack (SEFA) to boost transferability by randomly disrupting salient features. We provide theoretical analysis to demonstrate the superiority of the proposed method. In particular, perturbing the refined feature importance weighted intermediate features suppresses positive features and encourages negative features to realize adversarial attacks. Subsequently, self-ensemble is introduced to solve the optimization problem, thus enhancing the diversity from an optimization perspective. The diverse orthogonal initial perturbations disrupt these features stochastically, searching the space of transferable perturbations exhaustively to avoid poor local optima and improve transferability effectively. Extensive experiments show the effectiveness and superiority of the proposed SEFA, i.e., the success rates against undefended models and defense models are improved by 7.7% and 13.4%, respectively, compared with existing transferable attacks. Our code is available at https://github.com/chengshuyan/SEFA.

**Keywords** Black-box attacks · Transferability · Adversarial examples · Self-ensemble · Feature importance

## 1 Introduction

Breakthroughs in theories and technologies of deep learning have promoted the application of deep neural networks (DNNs) across diverse visual tasks, including autonomous driving [6, 7], medical diagnosis [4, 5], face recognition [2, 3], and image classification [1]. However, recent researches have revealed the vulnerability of DNNs against adversarial attacks, which add a carefully designed perturbation to the image to mislead DNNs. Researchers have proposed a range of defense methods for building safe and reliable deep-learning systems. Research on adversarial attacks is also vital for uncovering defects in DNNs and enhancing their robustness. In particular, researchers explore how different attack methods (e.g., adversarial sample attacks) can discover the vulnerabilities of models and how to enhance the robustness and resilience of models through defense techniques (e.g., adversarial training).

Based on the accessible information of target models, black-box and white-box attacks have developed into two main branches [19, 20]. White-box attacks allow attackers full access to target models and manipulate input images to optimize adversarial objectives. However, attackers are typically unable to access the complete information of a target model, which inspires studies on black-box attacks. Black-box attacks are grouped into query and transferable attacks based on whether a surrogate model is required. Query attacks [12, 13] refer to an attacker repeatedly querying a target model and using its feedback to generate adversar-

✉ Peng Li
lipeng@njupt.edu.cn

He Xu
xuhe@njupt.edu.cn

Yudong Yao
yyao@stevens.edu

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

[2] Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ial samples. Transferable attacks [10, 11, 16, 19] fool an unknown target model using an adversarial sample generated by a source model. Query methods compute the gradient while consuming excessive quantities of queries to update the adversarial perturbations/examples, which limits their practicality. Instead, relying on the adversarial transferability of examples across models, transferable attacks that utilize surrogate models instead of target models have received more attention.

Existing attacks [10, 14] suffer from limited transferability because adversarial examples tend to overfit to the source model. During attacks, adversarial samples may fall into a local optimum against a source model. Such local optimal solutions cannot be efficiently transferred across different models, thus limiting the availability of attacks. Strategies for improving transferability include data augmentation, gradient estimation, appropriate optimization objectives, ensemble models, and analysis of specific models. Some studies have addressed overfitting to surrogate networks through data enhancement, e.g., translation [11] and random transformation [16]. In terms of the gradient estimation, in several studies [10, 19, 29], attacks have been executed to solve the optimization problem using momentum [10], Nesterov [26], variance tuning [19], etc. In addition, in many studies [17, 18, 20, 23, 27], intermediate-layer based loss functions have been designed for higher transferability, which drives the development of feature-level attacks. Based on the findings of [28], some researchers have attempted to achieve more transferable adversarial attacks utilizing ensemble models [9, 28–30]. In a recent study [31], skip connections are used to generate transferable adversarial examples. Nevertheless, existing feature-level attacks disturb the features/attentions in a deterministic gradient-based manner. The generated perturbations minimize or maximize the given loss function in a relatively deterministic manner, which lacks diversity. Therefore, these attacks tend to fail to explore an abundant local optimum and thus suffer from limited transferability. In [45], DNNs are used to parameterize the adversary's generators for producing perturbations. These DNNs learn to produce adversarial perturbations using latent codes and randomly disrupt various prominent features for transferability. Attentive diversity attacks [45] rely on a generative adversarial network (GAN) that requires abundant training, and the GAN is another intricate and incomprehensible neural network.

Ensemble is one of the most powerful techniques for enhancing the DNN performance. Allen-Zhu et al. [47] investigate the working mechanism of ensemble in deep learning by considering the learning of multi-view data. This demonstrates that an ensemble of individual networks with random initialization can extract more comprehensive features. Based on the success of the ensemble method, Xu et al. [48] introduce self-ensemble. The fine-tuning of the model is improved via self-ensemble and self-distillation, and knowledge extraction is used to improve the fine-tuning efficiency of the ground truth and models. Self-distillation allows models to benefit from each other, while self-ensemble improves model performance by aggregating intermediate pre-trained models from different time points in the past as base models. In addition, in some studies, ensemble adversarial attacks are proposed for transferability. By treating iterative ensemble attacks as a gradient descent process, researchers [16] decrease the variance of gradients during the ensemble process to boost transferability. Xiong et al. [30] adopt two ensemble strategies and demonstrate that greater diversities in surrogate ensembles facilitate stronger transferability.

This study aims to enhance the transferability of adversarial perturbations. We propose a self-ensemble based feature-level adversarial attack (SEFA), which introduces diverse initializations. In particular, we designed a feature-level optimization objective with respect to the perturbations and sample the orthogonal initial perturbations as inputs to the optimization. Diverse orthogonal initializations guide the optimization process to search the perturbation space as much as possible to prevent adversarial perturbations from being trapped in model-specific local optima and to enhance the transferability of adversarial perturbations. To reduce the impact of model-specific information, refined aggregate gradients asfeature importances eliminate the noise caused by aggregation and directs the optimization objective to concentrate on the salient features corresponding to gradients with higher intensity, which guides the perturbation toward a more transferable direction.

Our contributions are summarized as follows.

- We introduce self-ensemble for increasing the randomness of the initial perturbations, i.e., sampling orthogonal low-dimensional vectors and fitting them to the perturbation space as the initial perturbations.
- We propose the self-ensemble based adversarial attack in feature space disrupting the salient and critical features distinguished by the refined aggregate gradient in a stochastic manner.
- We develop the combination of SEFA and other transferability enhancement methods, which generates more transferable adversarial perturbations and demonstrates the flexibility of the proposed method.

The remainder of this paper is organized as follows. In Section II, we provide a concise review of methods pertaining to adversarial attacks and defenses. Section III introduces the preliminaries. Section IV details the proposed SEFA. Section V presents empirical evaluations of SEFA and its comparisons with some baseline attacks. Finally, Section VI presents the conclusions of this research.

## 2 Related work

In this section, we introduce the literature related to adversarial attacks and adversarial defenses.

### 2.1 Adversarial attack

Szegedy et al. uncover the adversarial vulnerability of DNNs [15], stating that imperceptible perturbations manipulate the decisions of neural networks. They reveal two properties of the adversarial examples, misleading and imperceptible. Initially, some studies focus on adversarial attacks against image classification. Then, many attempts prove that the vulnerability exists in diverse mainstream visual tasks, e.g., face recognization [2, 3], smart healthcare [4, 5], automatic drive [6, 7]. The generation of adversarial examples has gradually developed into a technique, adversarial attacks. The transferability of adversarial perturbations seriously affects the development of deep learning, especially DNNs-based safety- and security-sensitive applications. Moreover, utilizing adversarial perturbations to boost the robustness of DNNs and investigate their defects gradually becomes a key issue. Therefore, adversarial attacks have received considerable attention among researchers.

Existing mainstream existence interpretations of adversarial issue focus on the linear nature of neural networks and non-robust features of datasets. One popular hypothesis on the adversarial examples is the high dimensional linear property [8], which also explains their generalization across datasets and models. [21] demonstrates that the adversarial nature of examples results from non-robust features, which are effective sources for achieving higher accuracy in neural networks and provide an explanation for transferability. The authors perform training and testing on both robust and non-robust features, highlighting that human have a limited perception of non-robust features, yet these features significantly influence decisions of the model.

Black-box attacks and white-box attacks are the two main branches of adversarial attacks. In white-box attacks, the attackers know the detailed information about the target model and can accurately generate adversarial perturbations via computing gradients of the optimization objective with respect to the images. However, in practice, attackers usually lack access to target models. In comparison, black-box attacks can work without model details, which is practical and challenging. Query attacks and transferable attacks evolve into two main categories of black-box attacks, where query attacks do not require surrogate models. Query attacks craft adversarial examples relying on the approximated gradients obtained by queries. Specially, a model assigns scores (i.e., soft label) to practicable labels for a given input and selects the label with the highest score as final decision (i.e., hard label). With soft labels, score-based attacks generate adversarial samples based on the responses of target models to fool those models. Instead, decision attacks solely require the hard labels, estimating gradients and updating adversarial examples to adjust the examples on the decision boundary. However, query-based attacks are impractical in real-world scenarios due to the large number of queries. More practical and flexible transfer-based attacks [9–11, 19–22], which rely on the transferability of adversarial perturbations, receive widespread attention. Adversarial samples are generated by surrogate models for attacking unknown target neural networks.

Many studies attempt to enhance the transferability of adversarial perturbations. [16] introduces random transformation (i.e., input diversity) and Dong et al. [11] demonstrate translation invariance of neural networks. The gradients are translated during iterations to enhance the robustness. Exploring the relationship of [16] and [11], [24] proposes resized-diverse-inputs (RDIM) and diversity-ensemble (DEM) further boosting the transferability of perturbations. They aggregate multi-scale gradients generated by RDIM with region fitting during iterations to generate transferable adversarial perturbations. [25] aggregates gradients of adversarial samples and their neighboring points during iterations to stabilize the oscillation of update directions, which boosts the transferability of adversarial samples to diverse target networks to a certain extent through data augmentation.

Moreover, improving optimization method for adversarial attacks is also a feasible direction, i.e., gradient ascent method. Momentum [10], Nesterov accelerated gradient [26] and variance tuning [19] are adopted during the iterations to find better local optima by avoiding oscillations in updates. Intuitively, a series of gradient descent methods can be used to address the optimization problem in adversarial attacks.

Aggregating surrogate networks improves the probability of finding transferable adversarial perturbations. In order to transfer the adversarial examples to unknown target networks, Liu et al. [28] aggregate the gradients of the surrogate models. They demonstrate that it is challenging for the target adversarial samples to transfer together with their target labels to other models, since the label distributions around the source labels differ among models with different architectures, even if the architectures are similar. Li et al. [29] fuse diverse feature-based models via vertical ensemble, devising ghost networks of the source model for transferability. Xiong et al. [30] demonstrate that there are variances between different model gradients in the aggregation process. They reduce the variance by stochastic variance reduction for stabilizing gradient update directions, making updated gradients more general to the other models. By contrast, Wu et al. [31] introduce decay parameters to reduce the gradients of residual modules during the computation of model gradients. With the skip connections similar to ResNet, they generate adversarial samples with higher transferability.

Based on the conclusion that different models share similar features, many attacks disturb the intermediate layers rather than the output layer, maximizing internal feature distortion for achieving higher transferability. Naseer et al. [17] maximize the distance of adversarial intermediate features and legitimate intermediate features, which pushes adversarial images away from the original images. Ganeshan et al. [18] utilize a discriminative criterion, namely the mean of channels, to guide the optimization. This method enhances the features that do not support the ground truth but suppressed those supporting the truth class to deceive the source network and target networks. However, it may get into a local optimization of a particular network. By contrast, feature importance-aware attack (FIA) [20] designs an appropriate optimization objective and aggregates gradient for eliminating model-specific information and generating adversarial perturbations. Lu et al. [27] delve into transferability across vision tasks and achieved powerful cross-task adversarial attacks by dispersion reduction. However, these attacks perturb examples along the gradients during the iterations, lacking of stochasticity, and therefore often trap into poor specific optima and exhibit limited transferability.

## 2.2 Adversarial defense

Adversarial examples can be used to investigate the internal shortcomings of DNNs and to improve their robustness. Researchers have proposed many adversarial defense approaches [32–41] to boost the robustness of neural networks, which are categorized into detection only and complete defense. The goal of complete defense is to make the outputs of the target model consistent with expectations. For example, a classification model correctly classifies adversarial samples. Moreover, detection only aims to detect potential adversarial examples for rejection.

Complete defense consists of two mainstream directions: modified training/input and modified networks. Many works improve model robustness by adversarial training. Ding et al. [32] propose consensus-based enhancement samples for adversarial defense. The intensity of the red, green and blue components of the image are exchanged to generate the enhancement samples. The original and consensus samples are used to train models. In the testing phase, the prediction results of the test and consensus samples are counted. The category corresponding to the maximum value above the threshold is the final classification result. If the maximum value is below the threshold, the test sample is discriminated as an adversarial sample. Lau et al. [33] propose joint spatial attack to generate adversarial perturbations against images and intermediate features. Then [33] uses mixup method to provide interpolated images for the attack to enhance the adversarial training. Yin et al. [34] demonstrate that the difference in feature distribution between the original and adversarial samples leads to a trade-off between accuracy and robustness. [34] utilizes a class-conditional discriminator to learn class-discriminative and attack-invariant features, i.e., to learn similar distributions of the original samples and various attack samples. The neural networks endeavor to learn domain invariant features to deceive the class-conditional discriminator. Liu et al. [35] show the vulnerability of adversarial training against transferable adversarial samples. [35] introduces linear robustness and approximates it with Jacobian norm. In addition, perturbation-based saliency map regularization is employed to enhance interpretability. Li et al. [36] analyze the problem that standard gradient regularization leads to inconsistency between model robustness and gradient saliency. Then a significant-based gradient regularization is proposed to reduce the performance gap by introducing gradient significance in the regularization training.

In terms of modifying input, researchers have proposed a variety of defense strategies. [37] proposes a new defense method to reconstruct legitimate samples using collaborative GANs to filter the perturbation noise in adversarial samples. The robustness of the model is improved by training an attacker model to generate adversarial samples and training a defender model to reconstruct the original samples. Zhang et al. [38] propose meta-invariant defense as an attack-independent defense method to achieve generalizable robustness against unknown adversarial attacks. Jia et al. [39] address the overfitting problem in fast adversarial training and propose a positive priori-guided adversarial initialization. Zhao et al. [40] emphasize the limitations of point-by-point adversarial sampling and introduce variational adversarial defense for more robust decision bounds. Niu et al. [41] experimentally demonstrate the correlation between perturbations and image pixels and the effectiveness of simultaneously eliminating perturbations in multiple frequency bands. [41] proposes to compress multiple frequency bands simultaneously to reduce the perturbation and the perturbation attachment space to purify the adversarial samples. Downsampling the lower frequency bands to disperse the perturbations and compressing the channel size in the higher frequency bands.

There are a several studies for detecting antagonistic samples. Nowroozi et al. [42] extract random features from the spreading layer of the source network as input to the target network. Then multiple target networks are trained to detect adversarial samples. [43] and [44] propose a multi-classifier architecture for image tampering detection and adversarial attack detection. Considering the security of the closed decision of one-class classification and the good performance of two-class classification, [43] and [44] combine one-class classification and two-class classification to improve the detection performance.

## 3 Preliminaries

Assuming a neural network for classification $F_{\psi} : x \mapsto c_s$, where $x$ represents the legitimate image, $c_s$ is the ground truth, and $\psi$ denotes the information of the neural network. The objective of non-target adversarial attacks is to create an adversarial perturbation $\delta$, which is carefully produced and results in misclassification on the target model (i.e., $F_{\psi}\left(x^{adv}\right) \neq c_s, x^{adv} = x + \delta$). In general, $\ell_p$-norm is used to restrict perturbations. We then formulate the generation of adversarial perturbations as follows.

$$\arg\max_{x^{adv}} L_{\psi}\left(x^{adv}, c_s\right), \text{s.t.} \left\|x - x^{adv}\right\|_p \leq \epsilon. \tag{1}$$

Function $L_{\psi}(\cdot, \cdot)$ calculates the distance between the predicted and true labels, $\epsilon$ constrains the intensity of the perturbations, and $p = 0, 2, \infty$.

Many attempts have been made to address the above adversarial optimization with the full information of $F_{\psi}$. However, the idea is unrealistic in applications. A viable approach is to optimize the adversarial examples on an accessible surrogate model $F_{\theta}$. The surrogate model $F_{\theta}$ and target model $F_{\psi}$ have different architectures and parameters but aligned outputs, and thus attackers produce transferable adversarial examples with $F_{\theta}$ for attacks. Feature-level attacks maximize distortions in intermediate features. $F_l(.)$ denotes the feature of an input from the $l$-th layer.

## 4 The proposed method

Following the observation that DNNs tend to extract similar features, feature-level attacks craft adversarial examples by perturbing the intermediate features. Therefore, adversarial examples generated in feature space have higher transferability, thus allowing them to fool multiple neural networks. However, these attacks generate perturbations in a specific gradient-based manner. During the iterations, they maximize a given loss function to update the perturbation. Owing to the lack of stochasticity in the process, they are often trapped in model-specific local optima, thereby reducing their transferability. Therefore, it is crucial to prevent the local optima to enhance the transferability. The production of adversarial perturbations requires fine-grained and model-agnostic features as guidance, i.e., feature importance, and updating adversarial examples requires diversity. To address these issues, we design a self-ensemble based feature-level adversarial attack. The devised approach significantly boosts the transferability of adversarial perturbations by refining the feature importance and introducing stochasticity into the optimization process, as exhibited in Fig. 1.

### 4.1 Self-ensemble for transferable adversarial attacks

In existing studies, adversarial attacks have been modeled as optimization problems and the adversarial example has been updated in a deterministic manner. Therefore, they often fall into local optima and suffer from limited transferability. In this study, we attempt to add diverse perturbations to clean images. Thus, obtaining an optimal $\delta$ can be formulated as the following constrained optimization problem,

$$\arg\min_{\delta} \|\delta\|_p, \text{s.t.} F_{\theta}(x + \delta) \neq c_s. \tag{2}$$

However, solving problem (2) is not trivial because it is impractical to determine a search space that satisfies the constraint. We obtain the perturbations by maximizing the loss function, as in the majority of previous studies (Fig. 2).

$$\arg\max_{\delta} L_{\theta}(x + \delta, c_s), \text{s.t.} \|\delta\|_p \leq \varepsilon, \tag{3}$$

where $L_{\theta}(\cdot, \cdot)$ is a loss function w.r.t. the perturbation $\delta$. Although problem (3) is not fully equivalent to (2) and may thus not guarantee that the obtained perturbations will always mislead the classifier, it quickly finds a possible perturbation within the constrained range.

The key to the problem (3) is an appropriate optimization objective. We design a new optimization objective for the problem (3) to perturb the object-aware features,

$$L(\delta) = \sum (W \odot F_l(x + \delta)), \tag{4}$$

where $W$ is the aggregate gradient (i.e., feature importance) w.r.t. $F_l(x)$.

$$W = \frac{\sum_{i=1}^{N} W_l^{x \odot B_{pd}^n}}{\left\| \sum_{i=1}^{N} W_l^{x \odot B_{pd}^n} \right\|_2}, \quad B_{pd}^n \sim \text{Bernoulli}(1 - p_d), \tag{5}$$

where the aggregate quantity $N$ is the number of masks for transform, and $p_d$ denotes the probability of random pixel dropping. $B_{pd}^n$, sampled from the Bernoulli distribution, Bernoulli (.), randomly discards the pixels of $x$ by the element-wise product with $x$. $W_l^x$ is expressed as follows,

$$W_l^x = \frac{\partial \ell(x, c_s)}{\partial F_l(x)}, \tag{6}$$

where $\ell(., .)$ donates an unnormalized probability w.r.t. the ground truth $c_s$. The sign of $W$ indicates the basic stance of the feature with respect to the truth class, and the intensity

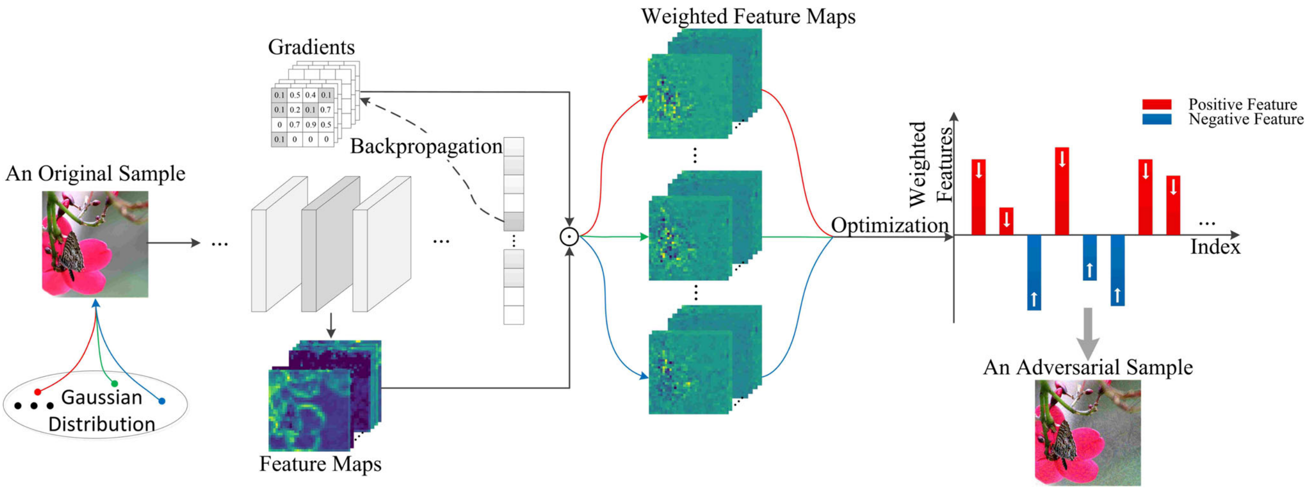


**Fig. 1** Overview of self-ensemble based feature-level adversarial attack. Given a clean image with an added random initial perturbation from a Gaussian distribution, intermediate feature maps are extracted from a surrogate network, and gradients are calculated as the feature importance backpropagating from the final probabilities to the feature maps. Then, the optimization of the weighted feature maps enhances negative features and suppresses positive features. The mutually orthogonal initial perturbations are selected successively from the Gaussian distribution to disrupt the features in a diverse manner, thus achieving higher transferability

denotes the importance of the feature. The features corresponding to positive and negative gradients are considered as positive and negative features of the samples, respectively. Minimizing $W \odot F_l(x + \delta)$ suppresses positive features but encourages negative ones, manipulating the learnable features of the samples. In contrast to previous attacks, i.e., feature disruptive attack (FDA) and neural representation distortion method (NRDM), $W \odot F_l(x + \delta)$ employs $W$ as a discriminator, thus focusing on salient object features and avoiding model-related information. Therefore, the weighted representation of the feature $W \odot F_l(x + \delta)$ is a powerful and transferable optimization objective that directs perturbations towards more transferable directions.

Similar to the training of neural networks, an optimization-based adversarial attack fixes the parameters of a pre-trained model and then optimizes the inputs to generate adversarial examples. The pre-trained model contains prior knowledge of the dataset distribution. The layers closer to the input layer capture fine-grained feature information common to multiple models with a high resolution through small receptive fields. However, the receptive fields and overlapping areas between them in the layers closer to the output layer gradu-

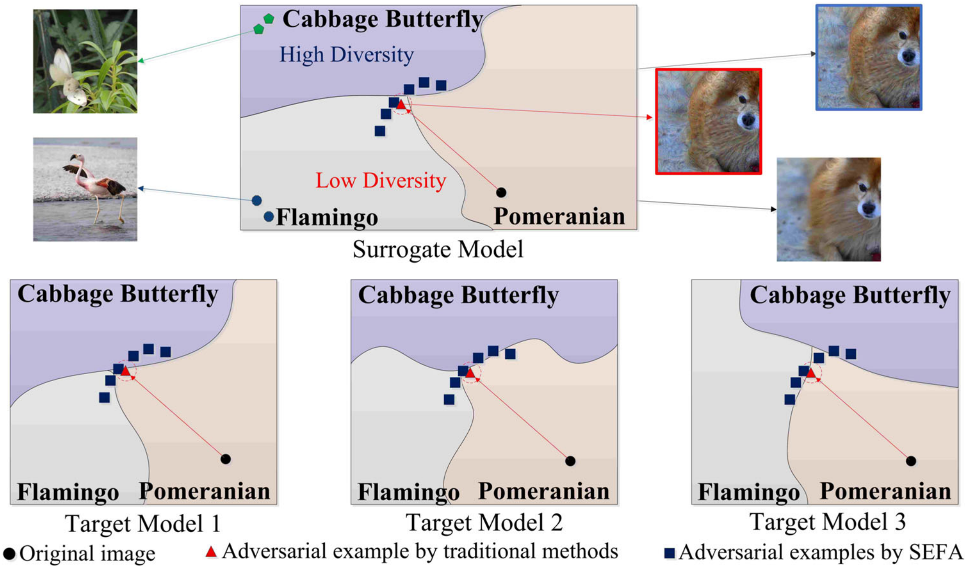


**Fig. 2** Illustration of adversarial examples produced by traditional attacks and the proposed SEFA among the class decision boundaries of the surrogate model and target models. Our attack attempts the feasible space of adversarial perturbations extensively then avoids the local optimum instead of greedily crafting deterministic adversarial examples like the traditional attacks that easily result in low local optima and then exhibit limited transferability among target models

ally increase, focusing on model-specific global information with rich semantic information. By observing the gradient of (4) w.r.t. $\delta$, we obtain

$$\frac{\partial L(\delta)}{\partial \delta} = \frac{\partial \sum (W \odot F_l(x+\delta))}{\partial F_l(x+\delta)} \frac{\partial F_l(x+\delta)}{\partial \delta}. \tag{7}$$

Updating the perturbation involves layers closer to the input layer, which indicates that the relatively general prior knowledge of the dataset, rather than the model-specific knowledge, is used to fine-tune and infect the source example.

In this study, we model adversarial attacks as the optimization problem (4). The optimization process significantly affects the transferability of adversarial perturbations. In addition, increasing diversity and variability can further improve the transferability. Therefore, we introduce diversity from an optimization perspective, sampling orthogonal initial perturbations in the outer layer of the generation and exploring as diverse directions as possible with a clean image as the origin. Diverse orthogonal initial perturbations guide the optimization to attempt different initial directions to increase the diversity and transferability of adversarial examples. In particular, given a trained DNN and an image, the constructed optimization objective requires an initial perturbation.

$$\delta_0 \sim \mathcal{N}\left(0.01, \sigma^2\right), \delta_0 \in \mathbb{R}^{Dim \times Dim}. \tag{8}$$

Because the dimension of the image $Dim$ is quite high, we sample the directions in the subspace and fit them to the original image space to improve the search efficiency, as shown in lines 4–7 of Algorithm 1.

$$\begin{aligned} \delta_0 \sim \mathcal{N}\left(0.01, \sigma^2\right), \delta_0 \in R^{\frac{Dim}{r} \times \frac{Dim}{r}}, \\ \delta_0 = \text{Interp}\left(\delta_0\right), \end{aligned} \tag{9}$$

where $r$ is the dimension reduction factor, and $\text{Interp}\,(.)$ denotes the bilinear-interpolation. In the $n$-th outer loop, the adversarial example $x$ is represented by $x_T^n$. Eventually, we obtain a set of adversarial examples $\left\{x_T^1, \cdots, x_T^{Num}\right\}$.

The use of SEFA can be understood as follows. Inspired by the ensemble, diverse adversarial examples are explored as much as possible to avoid falling into model-specific local optima, instead of moving the adversarial example along a deterministic gradient. The proposed SEFA embodies the idea of an ensemble without aggregation operations (e.g., averaging). We perform the attacks using the set $\left\{x_T^1, \cdots, x_T^{Num}\right\}$, preserving the diversity of the adversarial samples and improving the transferability of the attacks.

Comparing [30, 46], in which ensemble adversarial examples are generated by combining DNNs of different architectures, we explore adversarial samples with higher transferability generated by models of the same architecture but with randomness of initialization, which can be understood as a self-ensemble without external knowledge. Self-ensemble is utilized not only to enhance robustness [48], but also to generate adversarial examples. Xu et al. [48] aggregate intermediate pre-trained models of past time steps, while the proposed method constructs diverse adversarial examples as candidates by introducing randomness of initialization. Because we improve the transferability via an ensemble operation from an optimization perspective, it can be combined with previous transferability enhancement methods [9, 11, 16] to perform more powerful adversarial attacks.

**Algorithm 1** SEFA.

```
Require:
   The original image $x$, surrogate model $F$, intermediate layer $l$, filter probability $p_f$, drop probability $p_d$, ensemble number $N$, number of orthogonal perturbations $Num$, number of iterations $T$, dimension reduction factor $r$, bilinear interpolation Interp(.), and previous directions $S_{n-1} := \left\{\delta_0^j\right\}_{j=\max(n-1,1)}^{n-1}$.
Ensure:
   The set of adversarial images $\left\{x_T^1, \cdots, x_T^{Num}\right\}$.
 1: Initialize $\lambda = 1, \eta = \epsilon / T$;
 2: $W_f = \text{Filter}_{p_f}\left(\frac{\sum_{i=1}^{N} W_l^{x \odot B_{p_d}^n}}{\left\|\sum_{i=1}^{N} W_l^{x \odot B_{p_d}^n}\right\|_2}\right)$;
 3: for $n = 1$ to $Num$ do
 4:    $\delta_0^n \sim \mathcal{N}\left(0.01, \sigma^2\right), \delta_0^n \in R^{\frac{Dim}{r} \times \frac{Dim}{r}}$;
 5:    $\delta_0^n = proj_{span(S_{n-1})} \perp \delta_0^n$;
 6:    $S_n = S_{n-1} \cup \left\{\delta_0^n\right\}$;
 7:    $\delta_0^n = \text{Interp}\left(\delta_0^n\right)$;
 8:    $L\left(\delta_0^n\right) = \sum\left(W_f \odot F_l\left(x + \delta_0^n\right)\right)$;
 9:    $g_0 = 0$;
10:    for $t = 0$ to $T - 1$ do
11:       $g_{t+1} = \lambda g_t + \frac{\nabla_\delta L(\delta_t^n)}{\|\nabla_\delta L(\delta_t^n)\|_2}$;
12:       $\delta_{t+1}^n = \delta_t^n - \eta \cdot \text{Sign}\left(g_{t+1}\right)$;
13:       $x_{t+1}^n = \text{Clip}_{x,\epsilon}\left(x + \delta_{t+1}^n\right)$;
14:    end for
15: end for
16: return $\left\{x_T^1, \cdots, x_T^{Num}\right\}$;
```

## 4.2 Feature importance by refined gradient

The feature importance discussed in the previous section aggregates gradients from a randomly transformed $x$ to highlight robust/transferable features/gradients while neutralizing non-robust features or gradients. However, the aggregation of gradients introduces a small amount of noise, thus causing the optimization objective to focus on a few robust features from a limited set of transformed $x$, as shown in Fig. 3. The noise caused by aggregating gradients is inherently random and may not be shared by DNNs. Figure 3 illustrates the statistical information of the gradients of an

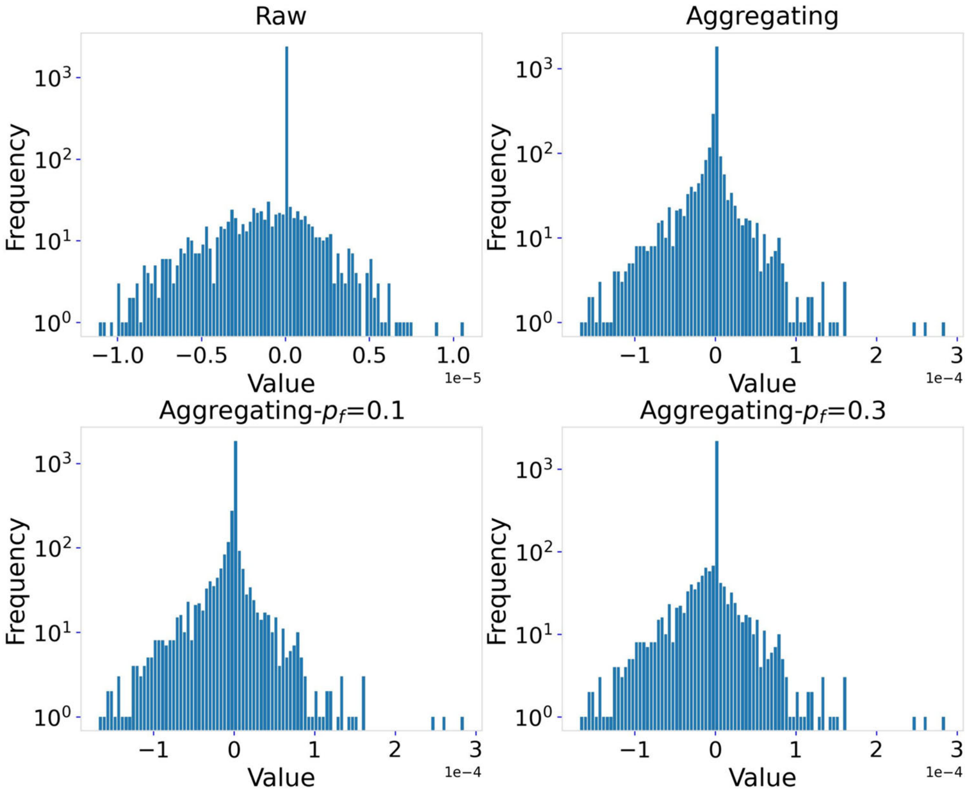


**Fig. 3** Histogram of the gradients. The gradients indicate the importance and stance of the feature w.r.t. the ground truth. The aggregation based on the raw gradient yields more values near zero. $p_f$ denotes the probability for filtering

image. The relatively flat distribution of the aggregated gradients indicates that the information of all the transformed images is observed, including random noise generated by the transformation.

To suppress the random noise, we propose the refining of the gradient, which refines the aggregate gradient from random transformations of the input image. The refinement operation eliminates redundancy while preserving the general texture and spatial structure. Object-aware and semantical salient features result in larger gradient magnitudes, which are further emphasized after aggregation. However, the gradients corresponding to random features have lower magnitudes. Refining retains the gradients corresponding to important features, while filtering out other gradients. In this study, we adopt quantile filtering with the probability $p_f$, which can be expressed as follows.

$$W_f = \text{Filter}_{p_f}\left(\frac{\sum_{i=1}^{N} W_l^{x\odot B_{p_d}^n}}{\left\|\sum_{i=1}^{N} W_l^{x\odot B_{p_d}^n}\right\|_2}\right), \qquad (10)$$

$$B_{p_d}^n \sim \text{Bernoulli}\,(1 - p_d)\,.$$

$\text{Filter}_{p_f}(.)$ is expressed as follows.

$$\text{Filter}_{p_f}(w) = \begin{cases} w_{i,j} & \left|w_{i,j}\right| > Q_{p_f} \\ 0 & \left|w_{i,j}\right| \leq Q_{p_f} \end{cases}, \qquad (11)$$

where $Q_{p_f}$ is the $100 \times p_f$-th percentile of $|w|$, $|.|$ denotes the absolute value of the input, and $w_{i,j}$ is a pixel at a cell $(i, j)$ of $w$.

The refined gradient $W_f$ preserves the highlighted critical and robust semantically meaningful features that provide more accurate information for generating transferable adversarial perturbations. Figure 4 presents visualizationes of the refined gradients. Compared to the aggregate gradient, the refined gradient is cleaner and focused on objects, thus providing better feature importance for transferable attacks. After the refinement operation, features with small absolute values are assigned to 0, and the contrast of the features is enhanced. In addition, the refined gradient focuses on the lower-left corner of the red box, i.e., the region of the maximum value of the features. The quantified filtering results are presented in Fig. 3. The change in the gradient distribution is more distinct when $p_f = 0.3$, whereas subsequent experiments show that $p_f = 0.1$ is more appropriate from the perspective of transferability.

### 4.3 Transferable adversarial attacks

By employing the refined gradient $W_f$ (i.e., the feature importance) and substituting (4) into (3), we obtain the following optimization objective for diverse perturbations in the

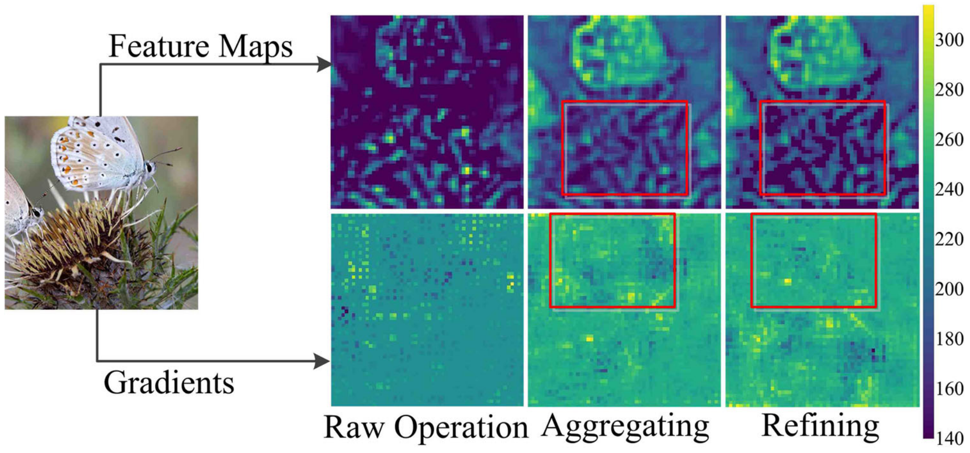


**Fig. 4** Visualization of feature maps and corresponding gradients at the layer Conv3_3 of VGG16. The raw operation provides the feature map and gradient from the clean image, the aggregate feature and gradient are calculated from multiple transforms of a legitimate image, and refining the aggregate feature and gradient generates the refined ones

feature space,

$$\arg\max_{\delta} \sum \left(W_f \odot F_l\left(x+\delta\right)\right), \text{s.t.} \|\delta\|_p \le \varepsilon. \tag{12}$$

The loss function in (12) enhances the salient features with a negative $W_f$ but suppresses those corresponding to a positive $W_f$. Thus, transferable adversarial attacks can be achieved.

**Algorithm 2** SEFA-DITI.

```
Require:
    The original image $x$, surrogate model $F$, intermediate layer $l$, filter probability $p_f$, drop probability $p_d$, ensemble number $N$, number of orthogonal perturbation $Num$, number of iterations $T$, previous directions $S_{n-1} := \left\{\delta_0^j\right\}_{j=\max(n-1,1)}^{n-1}$, dimension reduction factor $r$, bilinear-interpolation Interp(.), stochastic transform operation Trans $(x, p_T)$, Gaussian kernel generator Gkern (., .), and size of Gaussian kernel $Tkern\_size$.
Ensure:
    The set of adversarial images $\left\{x_T^1, \cdots, x_T^{Num}\right\}$.
Initialize $\lambda = 1$, $\eta = \epsilon / T$;
$W_f = \text{Filter}_{p_f}\left(\frac{\sum_{i=1}^N W_l^{x\odot B_{p_d}^n}}{\left\|\sum_{i=1}^N W_l^{x\odot B_{p_d}^n}\right\|_2}\right)$;
$Tkern = \text{Gkern}\left(Tkern\_size\right)$;
for $n = 1$ to $Num$ do
    $\delta_0^n \sim \mathcal{N}\left(0.01, \sigma^2\right), \delta_0^n \in R^{\frac{Dim}{r} \times \frac{Dim}{r}}$;
    $\delta_0^n = proj_{span(S_{n-1})^\perp} \delta_0^n$;
    $S_n = S_{n-1} \cup \left\{\delta_0^n\right\}$;
    $\delta_0^n = \text{Interp}\left(\delta_0^n\right)$;
    $L\left(x+\delta_0^n\right) = \sum\left(W_f \odot F_l\left(x+\delta_0^n\right)\right)$;
    $g_0 = 0$;
    for $t = 0$ to $T - 1$ do
        $g_{t+1} = \lambda g_t + \frac{\nabla_\delta L\left(\text{Trans}\left(x+\delta_0^n, p_T\right)\right)}{\left\|\nabla_\delta L\left(\text{Trans}\left(x+\delta_0^n, p_T\right)\right)\right\|_2}$;
        $\delta_{t+1}^n = \delta_t^n - \eta \cdot \text{Sign}\left(Tkern * g_{t+1}\right)$;
        $x_{t+1}^n = \text{Clip}_{x,\epsilon}\left(x + \delta_{t+1}^n\right)$;
    end for
end for
return $\left\{x_T^1, \cdots, x_T^{Num}\right\}$;
```

The key to transferability is introducing stochasticity in the generation of adversarial perturbations. Therefore, we propose the self-ensemble based transferable attack in the feature space. The effective adversarial attack framework comprises two subparts described in the previous subsections. The entire process is described in Algorithm 1. Given a clean image $x$, we generate $\delta_T^n$ through optimization with the initial perturbation $\delta_0^n$, a random variable from a Gaussian distribution. Subsequently, to make the perturbation as diverse as possible, we take the $\delta_0^{n+1}$ successively, which is orthogonal to the items of the perturbation set $\left\{\delta_0^0, \cdots, \delta_0^n\right\}$. Finally, we achieve diverse perturbations in the feature space, which boosts the transferability.

To further improve the transferability, we combine the proposed method with transferability enhancement approaches such as diverse input [11] and translation operations [16]. We describe the combination of methods in detail to explain this combination clearly. The combination of SEFA, diverse inputs iterative method (DIM) [11], and translation-invariant iterative method (TIM) [16] is referred to as SEFA-DITI, as shown in Algorithm 2. Trans (., .) is a random transform operation that creates diverse input images for transferability.

$$\text{Trans}(x, p_T) = \begin{cases} \text{Trans}(x) & \text{with probability } p_T \\ x & \text{with probability } 1 - p_T \end{cases}. \tag{13}$$

Gkern $(Tkern\_size)$ yields a two-dimensional Gaussian kernel $Tkern$ with $Tkern\_size$, which is used to convolve the gradient $g_{t+1}$.

Many previous gradient-based attacks have attempted to solve the optimization objective (12), such as the momentum iterative method (MIM) [10] and TIM [11]. Given the advantage and superiority of the momentum descent, the approach is used to address (12), as in [10], and algorithm 1 describes the details of the attack.

### 4.4 Theoretical analysis

Based on the finding that different neural networks extract similar features, NRDM [17] maximizes the distance between the features of the adversarial example $F_l\left(x^{adv}\right)$ and legitimate example $F_l\left(x\right)$. FDA [18] and FIA [20] perturb salient features carefully selected according to the activation values or gradients. For a better illustration, the objective functions are expressed as follows.

$$L_{NRDM}=\left\|F_l\left(x^{adv}\right)-F_l\left(x\right)\right\|_2, \tag{14}$$

$$\begin{aligned}L_{FDA}=&\log\left(\left\|F_l\left(x^{adv}\right)|F_l\left(x\right)<C_l\left(i,j\right)\right\|_2\right)\\&-\log\left(\left\|F_l\left(x^{adv}\right)|F_l\left(x\right)>C_l\left(i,j\right)\right\|_2\right),\end{aligned} \tag{15}$$

$$L_{FIA}=\sum\left(W\odot F_l\left(x^{adv}\right)\right), \tag{16}$$

where $C_l\left(i,j\right)$ denotes the mean activation values across channels.

With the optimization objectives above, FDA utilizes feature activation to characterize the importance of features, thus suppressing the features that support the ground truth but enhance the others. However, the distinguishable criterion, i.e., the mean activation values across channels, fails to effectively identify object-aware salient features and abate model-specific information. NRDM merely maximizes the distance $F_l\left(x^{adv}\right)$ and $F_l\left(x\right)$. In contrast, FIA achieves higher transferability by minimizing (16).

The gradients of the optimization objectives of NRDM (14), FDA (15), and FIA (16) are written as follows,

$$\frac{\partial L_{NRDM}}{\partial x^{adv}}=\frac{\partial\left\|F_l\left(x^{adv}\right)-F_l\left(x\right)\right\|_2}{\partial F_l\left(x^{adv}\right)}\frac{\partial F_l\left(x^{adv}\right)}{\partial x^{adv}}, \tag{17}$$

$$\begin{aligned}\frac{\partial L_{FDA}}{\partial x^{adv}}=&\frac{\partial\log\left(\left\|F_l\left(x^{adv}\right)|F_l\left(x\right)<C_l\left(i,j\right)\right\|_2\right)}{\partial F_l\left(x^{adv}\right)}\frac{\partial F_l\left(x^{adv}\right)}{\partial x^{adv}}\\&-\frac{\partial\log\left(\left\|F_l\left(x^{adv}\right)|F_l\left(x\right)>C_l\left(i,j\right)\right\|_2\right)}{\partial F_l\left(x^{adv}\right)}\frac{\partial F_l\left(x^{adv}\right)}{\partial x^{adv}},\end{aligned} \tag{18}$$

$$\frac{\partial L_{FIA}}{\partial x^{adv}}=\frac{\partial\sum\left(W\odot F_l\left(x^{adv}\right)\right)}{\partial F_l\left(x^{adv}\right)}\frac{\partial F_l\left(x^{adv}\right)}{\partial x^{adv}}. \tag{19}$$

The comparison of (17), (18), (19), and (7) as well as the experiments mentioned in the corresponding references, clearly indicates that the gradient of feature map w.r.t. the input $\frac{\partial F_l\left(x^{adv}\right)}{\partial x^{adv}}$ is the core, and the items (e.g., $\frac{\partial\left\|F_l\left(x^{adv}\right)-F_l\left(x\right)\right\|_2}{\partial F_l\left(x^{adv}\right)}$) introduce model-specific information, limiting transferability. While the item $\frac{\partial\sum\left(\Delta\odot F_l\left(x^{adv}\right)\right)}{\partial F_l\left(x^{adv}\right)}$ contains the feature importance guiding the adversarial example towards a more transferable direction. The conclusions and experiments in their respective papers adequately illustrate the advantages of FIA. However, it is difficult for FIA to introduce randomness from the perspective of perturbation. FIA only moves the adversarial sample along the deterministic gradient. Therefore, we introduce a self-ensemble into the optimization process of (4) to expand the search space for transferable perturbations. We consider the orthogonal initial perturbations $\delta_0^n$ $(1,2,\cdots,Num)$ and explore various directions at the beginning of the optimization to boost transferability. The gradient of the optimization objective (4) during the iterations can be expressed as follows,

$$\begin{aligned}\frac{\partial L\left(\delta\right)}{\partial\delta}&=\frac{\partial\sum\left(W_f\odot F_l\left(x+\delta\right)\right)}{\partial F_l\left(x+\delta\right)}\frac{\partial F_l\left(x+\delta\right)}{\partial\delta}\\&=\frac{\partial\sum\left(W_f\odot F_l\left(x+\delta\right)\right)}{\partial F_l\left(x+\delta\right)}\frac{\partial F_l}{\partial Z_l}\frac{\partial Z_l}{\partial F_{l-1}}\cdots\frac{\partial F_1}{\partial Z_1}\frac{\partial Z_1}{\partial\delta}\\&=\frac{\partial\sum\left(W_f\odot F_l\left(x+\delta\right)\right)}{\partial F_l\left(x+\delta\right)}\frac{\partial F_l}{\partial Z_l}W_l\cdots\frac{\partial F_1}{\partial Z_1}W_1,\end{aligned} \tag{20}$$

where $Z_l=W_lF_{l-1}+B_l$, $F_l=\sigma\left(Z_l\right)$, and $\sigma\left(\cdot\right)$ is the activation function. In the case of ReLU, the gradient is

$$\begin{aligned}\frac{\partial L\left(\delta\right)}{\partial\delta}=&\frac{\partial\sum\left(W_f\odot F_l\left(x+\delta\right)\right)}{\partial F_l\left(x+\delta\right)}\left(\sigma\left(Z_l\right)\odot\left(1-\sigma\left(Z_l\right)\right)\right)\\&W_l\left(\sigma\left(Z_{l-1}\right)\odot\left(1-\sigma\left(Z_{l-1}\right)\right)\right)W\times\cdots\times\\&\left(\sigma\left(Z_1\right)\odot\left(1-\sigma\left(Z_1\right)\right)\right)W_1.\end{aligned} \tag{21}$$

FIA updates the adversarial examples with deterministic gradients (19) because $W_l,\cdots,W_0$ are fixed, and $\sigma\left(Z_l\right),\cdots,\sigma\left(Z_1\right)$ are stable. As shown in Fig. 2, the proposed SEFA introduces self-ensemble by taking orthogonal initial perturbations, which introduces diversity from the perspective of gradient. Diverse initialization results in different $\sigma\left(Z_l\right),\cdots,\sigma\left(Z_1\right)$, contributing to the crafting of diverse transferable adversarial examples. The self-ensemble is the key to the proposed SEFA. As demonstrated in the following experiments, self-ensemble can significantly improve the transferability.

## 5 Experimental results

In this section, we describe the extensive experiments conducted to evaluate the effectiveness of SEFA. First, the

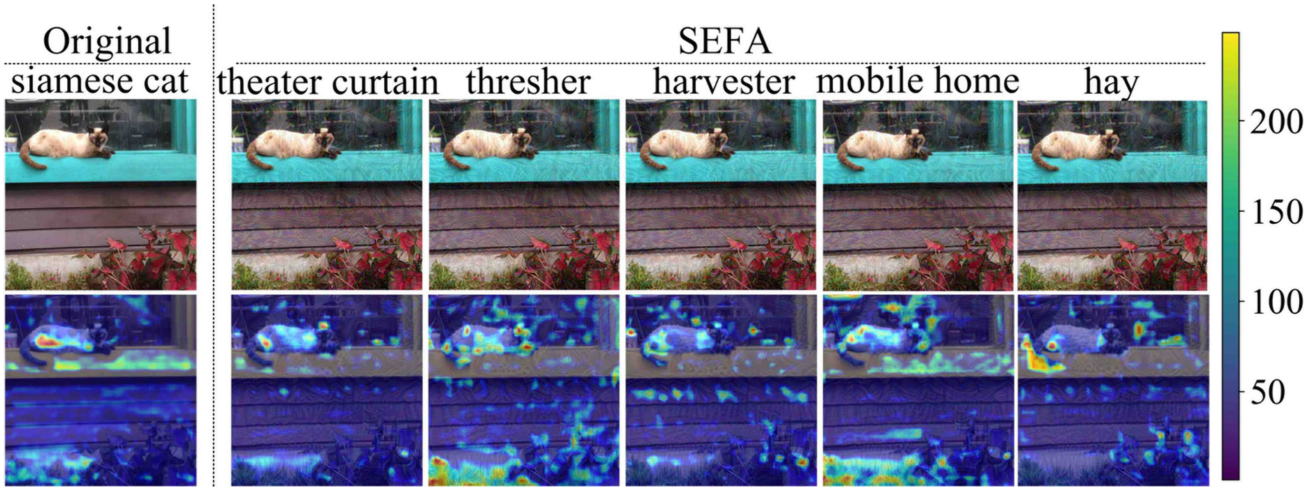


**Fig. 5** Legitimate/adversarial images (top row) produced by the proposed SEFA and their attentions (bottom row). SEFA generates adversarial examples that disrupt the attentions and final decisions in diverse manners

setup of the experiments is described. The attack results of SEFA against baseline methods with undefended models and advanced defended models are then illustrated. Furthermore, we perform ablation studies on the probability $p_f$ and hyper-parameter $Num$ in the proposed framework.

### 5.1 Experiment setup

Based on a baseline attack [20], we establish experimental settings to compare the transferability of adversarial attacks fairly. ImageNet [20] is a source dataset widely used for evaluating adversarial attacks. Figure 5 presents a legitimate image and some adversarial images. The perturbations generated by the proposed SEFA disrupt the attention and predictions of the model in various ways, thereby diversifying the enhancement of negative features and the suppression of positive features in the input images. The experiment setup is described as follows.

**Datatset.** The ImageNet-compatible dataset [20] is used to examine the transferability of the adversarial attacks, containing 1000 instances randomly sampled from different categories of the ILSVRC 2012 validation set. The CIFAR-10 dataset is a color image dataset containing 10 categories with 6000 images of size 32x32 in each category for training and evaluating image classification models.

**Models.** We test our method using four source models: ResNet-v1-152 (Res-152), Inception-ResNet-v2 (InceRes-v2), Inception-v3 (Ince-v3), and VGG16 (Vgg-16). Considering both normal and adversarial training, the proposed SEFA is used to attack several target models, seven normally trained models, and five defended models. For normal training, seven normally trained models are selected including Inception-ResNet-v2, ResNet-v1-50 (Res-50), ResNet-v1-152, VGG19 (Vgg-19), VGG16, Inception-v4 (Ince-v4), and Inception-v3. For the adversarial training [9], five adversarially trained models [1] are selected, namely, InceRes-v2-Ens, Ince-v3-Ens4, Ince-v3-Ens3, InceRes-v2-Adv, and Ince-v3-Adv. The source and target models above are pretrained with ImageNet, approaching an almost 100% classification success rate on the dataset. To evaluate the performance of the attack methods on the CIFAR10 dataset, we select four source models and seven target models for validation. Four source models, Res-20, Vgg-16, Shuffle-v2 [49], and RepVgg [50], are utilized to generate the adversarial samples. The seven target models are Res-20, Res-32, Vgg-16, Vgg-19, Mobile-v2 [51], Shuffle-v2, and RepVgg.

**Baseline Methods.** Several gradient iterative attacks are selected as baselines. In addition, three feature adversarial attacks, FIA [20], FDA [18], and NRDM [17], are selected as the comparison baselines. Our method is compared with these methods to validate the effectiveness and advancement of the proposed SEFA.

**Evaluation.** The probability that adversarial images generated by a source network mislead a target network is called the attack success rate. When the source network is the same as the target network, it indicates the success rate of the attack in a white setting. Instead, it is the black-box success rate of an attack.

**Parameter.** For a fair comparison, the parameters are set as follows, as in [20], step size $\eta = 1.6$, number of iterations $T = 10$, and perturbation limitation $\epsilon = 16$ . The momentum descent is a generic optimizer for all baselines, where the decay factor $\lambda$ is set as 1.0. For patch-wise attack method (PIM), the project kernel size $k_w$ is 3, project factor $\gamma$ is 0.5, and amplification factor $\beta$ is 2.5. The filter probability $p_f$

[1] https://github.com/tensorflow/models/tree/archive/research/adv_imagenet_models

and $Num$ in SEFA are 0.1 and 50, respectively, and $r$ is 2. In SEFA-DITI, the kernel size $Tkern\_size$ is 15, and the transform probability $p_T$ is 0.7. As for the target layer of the surrogate model in feature-level attacks, we select the middle layer of the networks for attacks. Mixed_5b in Ince-v3, the last layer of block2 in Res-152, Conv_4a in InceRes-v2, and Conv3_3 in Vgg-16 are selected as the target layers. Under these settings, we could realize a fair comparison between the devised SEFA and baseline attacks.

## 5.2 Comparison of transferability

This section presents the performance of the baseline attacks and proposed SEFA against normally trained models and adversarially trained ones respectively. We choose four source models with different architectures (Ince-v3, InceRes-v2, Res-152, and Vgg-16) and attackdefense models and normally trained models.

**Attacking Normally Trained Models.** The proposed SEFA significantly outperforms the baseline attacks, as shown in Table 1. The success rates of the adversarial examples produced by the source models Ince-v3 and InceRes-v2 increased by about 10%. In particular, the adversarial examples generated by SEFA with the surrogate network Vgg-16 successfully transfer among different models, achieving attack success rates of over 95%. The success rates of the combination of SEFA and DITI, i.e., SEFA-DITI, increase by 1% ∼ 3%. Comparing DITI, SEFA, and SEFA-DITI, it

**Table 1** Attack success rates of various attacks against normally trained models

| | Attack | Ince-v3 | Ince-v4 | InceRes-v2 | Res-50 | Res-152 | Vgg-16 | Vgg-19 |
|---|---|---|---|---|---|---|---|---|
| Ince-v3 | MIM | **100.0%** | 41.5% | 38.9% | 35.1% | 29.9% | 39.9% | 40.1% |
| | PIM | 97.8% | 55.2% | 50.9% | 53.1% | 46.0% | 61.6% | 61.8% |
| | DITI | 96.0% | 51.7% | 39.6% | 57.1% | 50.8% | 67.2% | 65.6% |
| | NRDM | 91.0% | 61.1% | 52.4% | 42.2% | 31.7% | 40.5% | 40.7% |
| | FDA | 81.7% | 42.7% | 36.3% | 30.1% | 25.0% | 32.0% | 31.5% |
| | FIA | 98.1% | 82.8% | 79.3% | 70.0% | 64.6% | 70.0% | 71.5% |
| | SEFA | 98.4% | ***86.7%*** | ***85.9%*** | ***81.3%*** | ***77.4%*** | ***80.2%*** | ***80.4%*** |
| | SEFA-DITI | ***98.7%*** | **88.3**% | **87.1**% | **83.7**% | **78.6**% | **83.4**% | **83.5** |
| InceRes-v2 | MIM | 60.1% | 52.0% | ***99.1%*** | 41.2% | 36.2% | 43.2% | 39.9% |
| | PIM | 65.0% | 62.8% | **99.5%** | 55.6% | 50.4% | 63.6% | 63.2% |
| | DITI | 60.7% | 57.7% | 89.9% | 61.8% | 58.9% | 69.8% | 66.3% |
| | NRDM | 72.6% | 67.2% | 76.4% | 55.4% | 45.3% | 50.5% | 51.7% |
| | FDA | 69.1% | 67.6% | 78.2% | 51.7% | 40.1% | 49.0% | 44.8% |
| | FIA | 80.0% | 76.3% | 88.3% | 71.7% | 68.5% | 71.2% | 71.4% |
| | SEFA | ***80.6%*** | ***77.7%*** | 89.7% | ***76.6%*** | ***73.5%*** | ***76.8%*** | ***75.9%*** |
| | SEFA-DITI | **83.7%** | **79.0%** | 91.8% | **78.4%** | **74.9%** | **78.7%** | **78.1%** |
| Res-152 | MIM | 57.1% | 48.9% | 42.5% | 90.5% | ***99.6%*** | 71.6% | 71.8% |
| | PIM | 65.9% | 56.7% | 51.3% | 92.7% | **99.9%** | 83.1% | 82.5% |
| | DITI | 54.3% | 48.2% | 41.5% | 82.9% | 98.9% | 78.5% | 76.6% |
| | NRDM | 64.0% | 59.3% | 50.7% | 87.8% | 95.6% | 79.2% | 79.3% |
| | FDA | 60.8% | 52.6% | 47.9% | 84.8% | 94.3% | 76.3% | 76.0% |
| | FIA | 85.1% | 81.0% | 76.8% | 97.1% | 99.3% | 91.2% | 91.1% |
| | SEFA | ***93.6%*** | ***90.2%*** | ***89.8%*** | ***98.4%*** | **99.9%** | ***94.6%*** | ***94.0%*** |
| | SEFA-DITI | **95.1%** | **92.7%** | **92.1%** | **98.9%** | **99.9%** | **95.7%** | **96.2%** |
| Vgg-16 | MIM | 82.3% | 82.6% | 78.6% | 90.2% | 85.5% | ***99.8%*** | 95.4% |
| | PIM | 84.2% | 82.1% | 75.7% | 91.1% | 85.6% | **100.0%** | 97.4% |
| | DITI | 80.8% | 81.7% | 68.1% | 80.2% | 78.2% | 94.5% | 95.2% |
| | NRDM | 73.1% | 72.2% | 57.7% | 78.1% | 74.1% | 94.5% | 91.8% |
| | FDA | 76.9% | 76.8% | 64.5% | 80.2% | 78.3% | 94.5% | 95.2% |
| | FIA | 94.9% | 95.7% | 91.7% | 97.2% | 94.9% | 99.5% | 99.6% |
| | SEFA | ***98.0%*** | ***97.8%*** | ***95.9%*** | ***98.5%*** | ***97.7%*** | **100.0%** | ***99.8%*** |
| | SEFA-DITI | **98.9%** | **98.5%** | **97.1%** | **99.3%** | **98.6%** | **100.0%** | **99.9%** |

The first column presents the source models (Ince-v3, InceRes-v2, Res-152, and Vgg-16). The best results are indicated in bold

**Table 2** Attack success rates of various attacks against normally trained models on the CIFAR10 dataset

| | Attack | Res-20 | Res-32 | Vgg-16 | Vgg-19 | Mobile-v2 | Shuffle-v2 | RepVgg |
|---|---|---|---|---|---|---|---|---|
| Res-20 | MIM | 93.8% | 65.3% | 44.0% | 44.6% | 64.4% | 54.2% | 60.0% |
| | PIM | 91.6% | 56.3% | 37.2% | 36.7% | 58.9% | 49.9% | 50.8% |
| | DITI | 82.4% | 46.0% | 31.2% | 31.9% | 50.1% | 48.4% | 43.8% |
| | FIA | **99.4%** | ***77.2%*** | ***54.0%*** | ***53.0%*** | ***76.6%*** | ***63.5%*** | ***71.9%*** |
| | SEFA | ***98.9%*** | **82.2%** | **64.0%** | **62.8%** | **81.7%** | **70.8%** | **77.0%** |
| Vgg-16 | MIM | 59.7% | 58.4% | 76.5% | 58.2% | 59.5% | 52.8% | 58.6% |
| | PIM | 49.0% | 47.1% | 64.1% | 39.6% | 49.7% | 44.5% | 44.5% |
| | DITI | 43.0% | 40.7% | 68.3% | 39.8% | 45.3% | 45.8% | 41.8% |
| | FIA | ***70.1%*** | ***69.6%*** | ***83.7%*** | ***70.1%*** | ***70.2%*** | **64.2%** | ***69.8%*** |
| | SEFA | **70.8%** | **71.3%** | **88.7%** | **71.5%** | **71.2%** | ***63.0%*** | **70.7%** |
| Shuffle-v2 | MIM | 55.8% | 51.7% | ***39.6%*** | ***38.8%*** | 60.4% | 93.4% | 49.7% |
| | PIM | 48.9% | 45.0% | 32.5% | 32.1% | 53.2% | 88.0% | 42.5% |
| | DITI | 37.3% | 33.4% | 25.6% | 24.9% | 42.1% | 80.2% | 32.9% |
| | FIA | ***57.5%*** | ***52.6%*** | 38.0% | 38.0% | ***62.6%*** | **97.4%** | ***49.8%*** |
| | SEFA | **61.4%** | **56.9%** | **43.9%** | **42.8%** | **66.6%** | ***96.6%*** | **53.8%** |
| RepVgg | MIM | 54.4% | 53.0% | 42.6% | 42.2% | 55.1% | 47.2% | 68.5% |
| | PIM | 51.1% | 48.7% | 35.3% | 35.6% | 52.7% | 44.5% | 74.3% |
| | DITI | 40.6% | 38.0% | 29.1% | 29.7% | 42.8% | 41.6% | 60.8% |
| | FIA | ***76.5%*** | ***75.7%*** | **64.3%** | **64.1%** | ***76.4%*** | **66.7%** | ***92.5%*** |
| | SEFA | **78.5%** | **77.0%** | ***60.5%*** | ***59.6%*** | **78.6%** | ***64.5%*** | **97.7%** |

The first column presents the source models (Ince-v3, InceRes-v2, Res-152 and Vgg-16). The best results are indicated in bold

can be observed that SEFA contributes more prominently to SEFA-DITI than DITI. With the source models Inception-ResNet-v2 and Inception-v3, SEFA has higher success rates for both black-box and white-box attacks than existing feature-based attacks such as FIA, FDA, and NRDM. Compared to previous studies, SEFA improves the success rate against normally trained models by about 7.7%.

The adversarial examples generated by the surrogate network Vgg-16 mislead the target models with success rates of nearly 96%, while the transferability of the adversarial perturbations with surrogate networks Ince-v3 and InceRes-v2 is limited. The results in Table 1 imply that less complicated models (e.g., Vgg-16) tend to craft more transferable adversarial examples because these models avoid examples overfitting to the source models compared to complex/large ones (e.g., Ince-v3 and InceRes-v2). It would be interesting to explore more appropriate models for generating transferable adversarial perturbations.

For the CIFAR10 dataset, the results of the attacks against the seven target models are exhibited in Table 2. The layer3_2 of Res-20, features_40 of Vgg-16, stage4_2 of Shuffle-v2, and stage4_0 of RepVgg are selected as the target intermediate layers for FIA and SEFA. Table 2 shows that the proposed SEFA obtains better results. For example, SEFA achieves an average improvement of 2.7 % in the attack success rate compared to FIA. Thus, SEFA is effective for both the large-scale ImageNet dataset and simple CIFAR10 dataset with a wide applicability.

We conduct experiments against feature randomization (FR) [42] on the CIFAR10 dataset to verify the effectiveness of the attack methods. Following [42], we determine the feature size $FS = \{30, 50, 200, 400, NS\}$. $NS$ denotes the full size of the flatten layer of the source model Res-20. There are 50,000 original samples and 50,000 adversarial samples for training, 5,000 original samples and 5,000 adversarial samples for validation, and 5,000 original samples and 5,000 adversarial samples for testing. First, we input the samples into the source model to extract features. Subsequently, we randomly select 50 times feature vectors from the features. The Fifty sets of feature vectors of size $fs \in FS$ are used to train 50 support vector machines (SVMs). In the evaluation

**Table 3** Attack success rates of various attacks in mis-match index testing against FR on the CIFAR10 dataset

| Attack | 30 | 50 | 200 | 400 | NS |
|---|---|---|---|---|---|
| MIM | 71.07% | 63.05% | 28.86% | 48.57% | 99.36% |
| PIM | 70.49% | 62.65% | 28.73% | 47.91% | 99.83% |
| DITI | 70.55% | 62.61% | 28.07% | 48.1 % | 99.66% |
| FIA | 71.10% | 64.03% | 29.16% | 48.64% | 99.65% |
| SEFA | 72.21% | 64.14% | 29.96% | 49.24% | 99.54% |

phase, we use the 50 SVMs to identify the 50 feature sets of adversarial samples for different attacks. The success rates of multiple attacks against FR in mis-match index testing are shown in Table 3. The experimental results demonstrate a slight improvement in the attack performance of the proposed SEFA against FR. For example, the attack success rate of SEFA improves by an average of 0.5% over FIA. FR is an effective adversarial detection method. The robustness of attack methods against defenses needs to be improved to further facilitate defenses.

**Attacking Defense Models.** Adversarial training of neural networks has regularization-like effects, achieving strong robustness to adversarial examples. In most cases, the proposed SEFA and SEFA-DITI are ranked among the top two, as shown in Table 4. This is because SEFA-DITI combines SEFA and the enhancement methods DIM and TIM to introduce randomization. Data enhancement helps to further improve the generalization of the adversarial samples and hence the transferability. Compared to the baseline attack, our approach improves the success rate against the defense model by about 13.4%. Compared with normally trained models, the proposed SEFA demonstrates a more significant improvement in attacking the adversarially trained networks. This is because the success rates of attacks against the normal training models are already quite high. There is a small number of difficult samples for attackers. Therefore, it is difficult to improve the success rates. Table 4 demonstrates the threats posed by the proposed SEFA to the defense models.

**Table 4** Attack success rates of various attacks against defense models

| | Attack | Ince-v3-Adv | Ince-v3-Ens3 | Ince-v3-Ens4 | InceRes-v2-Adv | InceRes-v2-Ens |
|---|---|---|---|---|---|---|
| Ince-v3 | MIM | 20.8% | 15.1% | 15.3% | 16.4% | 6.8% |
| | PIM | 34.1% | 32.3% | 38.1% | 30.8% | 26.0% |
| | DITI | 43.2% | 41.8% | 46.5% | 37.2% | 33.0% |
| | NRDM | 27.5% | 8.6% | 12.3% | 19.2% | 5.0% |
| | FDA | 19.5% | 8.9% | 12.0% | 12.2% | 5.0% |
| | FIA | 53.6% | 43.2% | 41.1% | 53.5% | 22.8% |
| | SEFA | ***66.1%*** | ***56.6%*** | ***56.5%*** | ***67.3%*** | ***34.9%*** |
| | SEFA-DITI | **68.2%** | **57.9%** | **58.1%** | **68.7%** | **37.2%** |
| InceRes-v2 | MIM | 26.5% | 15.4% | 16.8% | 23.3% | 9.8% |
| | PIM | 39.1% | 38.9% | 41.3% | 34.5% | 32.1% |
| | DITI | 51.7% | 52.3% | 54.4% | 55.7% | ***50.2%*** |
| | NRDM | 36.6% | 15.8% | 16.6% | 29.0% | 7.9% |
| | FDA | 34.2% | 16.2% | 15.7% | 29.9% | 7.9% |
| | FIA | 54.9% | 45.3% | 43.7% | 55.5% | 36.3% |
| | SEFA | ***65.3%*** | ***60.3%*** | ***57.3%*** | ***66.9%*** | 49.0% |
| | SEFA-DITI | **67.2%** | **62.5%** | **59.4%** | **68.9%** | **51.3%** |
| Res-152 | MIM | 41.1% | 41.8% | 42.3% | 39.9% | 22.0% |
| | PIM | 50.1% | 51.5% | 50.5% | 46.3% | 38.5% |
| | DITI | 52.6% | 55.4% | 61.7% | 49.9% | 48.2% |
| | NRDM | 57.1% | 48.0% | 45.9% | 42.2% | 36.5% |
| | FDA | 56.5% | 43.9% | 40.9% | 40.6% | 35.5% |
| | FIA | 70.0% | 61.4% | 59.3% | 66.1% | 41.5% |
| | SEFA | ***88.1%*** | ***84.9%*** | ***84.7%*** | ***83.3%*** | ***77.1%*** |
| | SEFA-DITI | **89.7%** | **85.8%** | **85.4%** | **84.5%** | **78.6%** |
| Vgg-16 | MIM | 65.3% | 67.3% | 67.7% | 64.2% | 46.2% |
| | PIM | 52.0% | 50.0% | 56.8% | 43.7% | 39.5% |
| | DITI | 72.7% | 67.1% | 68.6% | 60.6% | 57.6% |
| | NRDM | 69.3% | 67.2% | 67.9% | 59.5% | 56.7% |
| | FDA | 72.8% | 67.1% | 68.8% | 60.6% | 57.5% |
| | FIA | 87.5% | 86.2% | 85.0% | 86.7% | 70.7% |
| | SEFA | ***96.2%*** | ***95.2%*** | ***95.0%*** | ***94.0%*** | ***91.8%*** |
| | SEFA-DITI | **97.5%** | **96.7%** | **96.3%** | **95.4%** | **93.1%** |

The first column presents the source models (Ince-v3, InceRes-v2, Res-152 and Vgg-16). The best results are indicated in bold

Tables 1 and 4 present the attack success rates against the normally trained models and the adversarially trained models, respectively. The values in the tables indicate the attack success rates (corresponding to rows) against the target models (corresponding to columns).

### 5.3 Ablation study

There are two parameters in the proposed method: the filter probability $p_f$ and number of initial perturbations $Num$. With the parameter settings $p_f = 0.1$ and $Num = 50$, we fix one parameter and modify the other to analyze the effect of the parameters on the framework.

$p_f$ increases from 0 to 0.4 in steps of 0.1, and $Num$ increases from 0 to 70. Figures 6 and 7 illustrate the effects of $p_f$ and $Num$ on attacks. The effects of the filtering probability and quantity of initial perturbations on the success rates of the source and target networks are approximately the same. The trends in the attack success rates for different target networks with $Num$ increasing are also approximately consistent. The attack time increases as $Num$ increases, and the attack success rate gradually became saturated. Therefore, the optimal $Num$ for attacking is 50 to achieve a better tradeoff between effectiveness and efficiency, as shown in Fig. 6. In terms of the filter probability, a larger $p_f$ (e.g.,

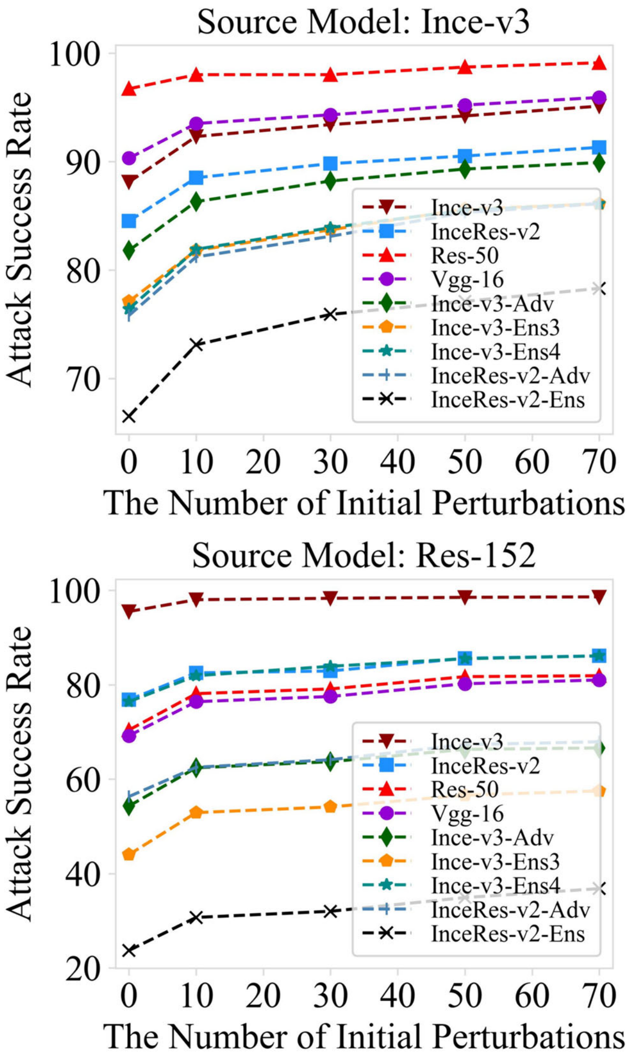


**Fig. 6** Effects of the number of initial perturbations on the attack success rate. Two source models, Ince-v3 and Res-152, generate adversarial examples with different filter probabilities. The filter probability changes from 0 to 70. The success rates are the results of attacking four normally trained models Ince-v3, Vgg-16, InceRes-v2, and Res-152 and five defense models InceRes-v2-Adv, InceRes-v2-Ens, Ince-v3-Ens4, Ince-v3-Ens3, and Ince-v3-Adv

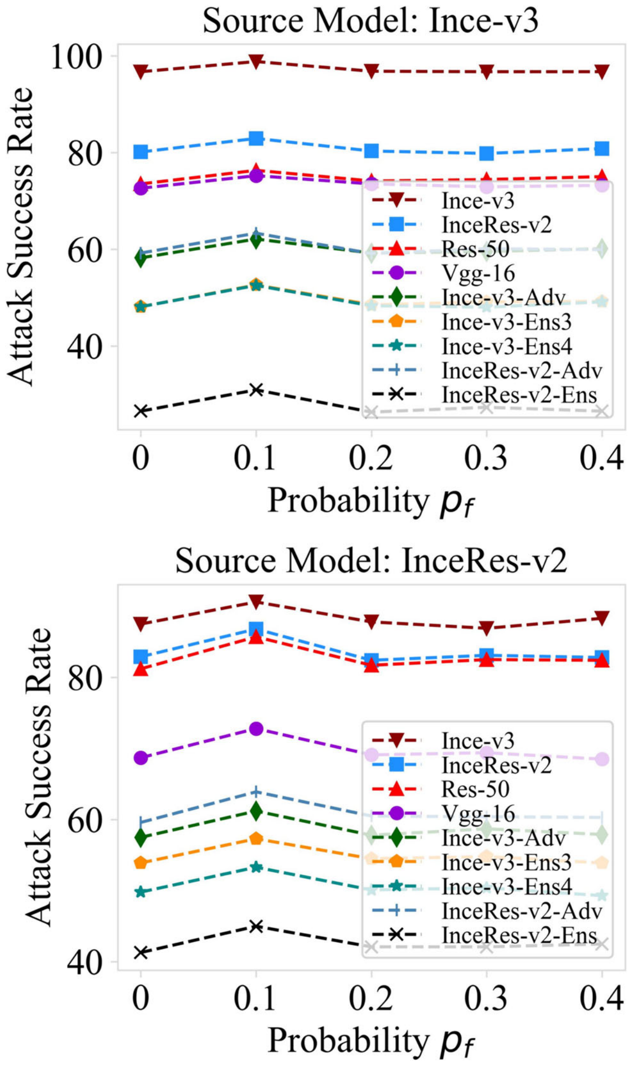


**Fig. 7** Effects of the filter probability on the attack success rate. Two source models, Ince-v3 and Res-152, generate adversarial examples with different filter probabilities. The filter probability changes from 0 to 0.4. The adversarial examples are used to attack four normally trained models Ince-v3, Vgg-16, InceRes-v2, and Res-152 and five defense models InceRes-v2-Adv, InceRes-v2-Ens, Ince-v3-Ens4, Ince-v3-Ens3, and Ince-v3-Adv

0.4) removes a large amount of redundant feature importance information. However, the success rates of attacks with $p_f = 0.1$ increases significantly, as shown in Fig. 7. Finally, the appropriate number of initial perturbations $Num$ and filter probability $p_f$ are selected for attack.

Moreover, the keys to the proposed SEFA are the stochasticity and refined gradient. To investigate the contributions of these two factors, we designed four optimization objectives and experimentally validated them using two source models: Ince-v3 and Res-152. The four objective functions are constructed as follows. $L_1$ boosts the positive features and discourages the negative features from the aggregate gradient, and $L_2$ uses a refined gradient. $L_4$ is the proposed loss function (4), and $L_3$ selects the aggregate gradient. Here, $L_2$ and $L_3$ explore the effects of these two items, respectively. Figure 8 presents the success rate with the four loss functions.

$$L_1 = \sum \left( W \odot F_l \left( x^{adv} \right) \right), \tag{22}$$

$$L_2 = \sum \left( W_f \odot F_l \left( x^{adv} \right) \right), \tag{23}$$

$$L_3 = \sum \left( W \odot F_l \left( x + \delta \right) \right), \tag{24}$$

$$L_4 = \sum \left( W_f \odot F_l \left( x + \delta \right) \right). \tag{25}$$

$L_2$ and $L_3$ outperform $L_1$, demonstrating the effectiveness of the two items-the refined gradient and stochasticity introduced. $L_3$ surpasses $L_2$, indicating that the stochasticity improves the transferability to a greater extent. In most cases, the proposed loss $L_4$ significantly outperforms the others, demonstrating the advantage of the proposed SEFA.

## 6 Conclusions

We propose a general framework for adversarial attacks by introducing self-ensemble. Our method disrupts the salient features in a stochastic manner through diverse initial perturbations and refining the feature importance, thus significantly improving the diversity and randomness of adversarial perturbations. Consequently, the generated adversarial examples efficiently avoid being trapped in model-specific local optima and become more transferable among the target models.

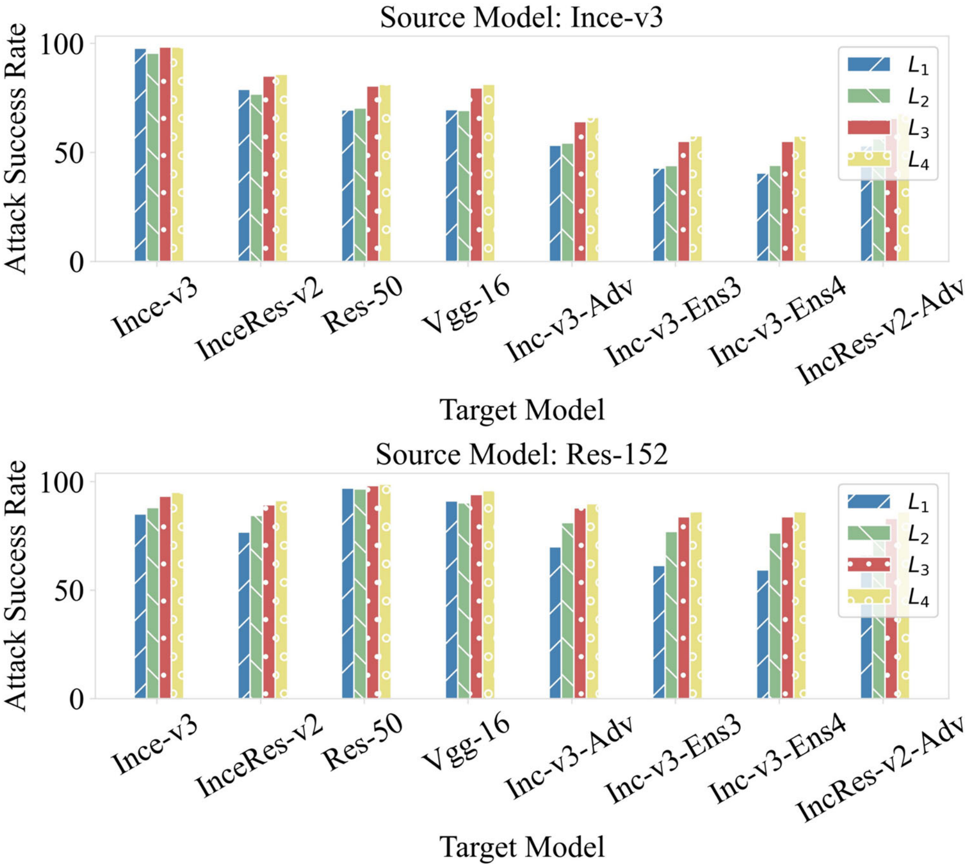


**Fig. 8** Effect of stochasticity and refined gradient. $L_1$ acts as the baseline. $L_2$ and $L_3$ comprise the stochasticity and refined gradient, respectively. $L_4$ adopts the above two terms simultaneously

Moreover, the devised attack can further enhance the transferability in combination with other methods. Theoretical analysis and extensive experiments demonstrate the excellent performance of SEFA against the baseline attacks.

In the future, we will consider reducing the computational complexity of the proposed method. Examining more effective methods based on self-ensemble, such as transferable targeted adversarial attacks, is a possible future research direction. Moreover, we intend to explore the generalization of adversarial attacks to various visual applications such as object detection and semantic segmentation.

**Acknowledgements** This study is sponsored by the National Natural Science Foundation of P. R. China (Nos. 61902196, 62102194, and 62102196) and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (No. KYCX22_1019).

**Author Contributions** Conceptualization: Shuyan Cheng;
Formal Analysis: Shuyan Cheng;
Methodology: Shuyan Cheng, Peng Li;
Investigation: Shuyan Cheng, Peng Li;
Resources: Shuyan Cheng;
Software: Shuyan Cheng, Jianguo Liu;
Supervision: Peng Li, He Xu, Yudong Yao;
Validation: Shuyan Cheng, Jianguo Liu;
Visualization: Shuyan Cheng, Jianguo Liu;
Writing - original draft: Shuyan Cheng, Peng Li, Yudong Yao;
Writing - review & editing: Shuyan Cheng, Peng Li, Yudong Yao;

**Data Availability** The datasets used in this study are publicly available, and relevant references have been included. The authors do not own any of the datasets used in this study.

## Declarations

**Conflicts of interest** The authors declare that they have no competing interests.

**Ethics approval** All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

## References

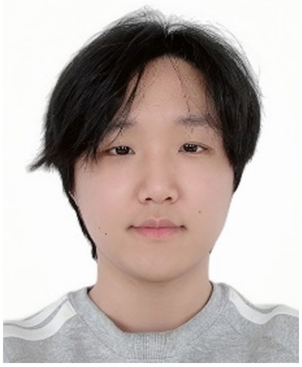
1. Xianfeng Ou, Meng Wu, Bing Tu et al (2023) Multi-Objective Unsupervised Band Selection Method for Hyperspectral Images Classification. IEEE Trans Image Process 32:1952–1965
2. Han Ruidong, Wang Xiaofeng, Bai Ningning et al (2023) FCD-Net: Learning to Detect Multiple Types of Homologous Deepfake Face Images. IEEE Trans Inf Forensics Secur 18:2653–2666
3. Lv Xianwei, Chen Yu, Jin Hai et al (2022) HQ2CL: A High-Quality Class Center Learning System for Deep Face Recognition. IEEE Trans Image Process 31:5359–5370
4. Kim Sekeun, Jiang Zhenxiang, Zambrano Byron A et al (2023) Deep Learning on Multiphysical Features and Hemodynamic Modeling for Abdominal Aortic Aneurysm Growth Prediction. IEEE Trans Med Imaging 42(1):196–208
5. Shu Yucheng, Li Hengbo, Xiao Bin, Bi Xiuli et al (2023) Cross-Mix Monitoring for Medical Image Segmentation With Limited Supervision. IEEE Trans Multimedia 25:1700–1712
6. Fan Bin, Yang Yuzhu, Feng Wensen et al (2023) Seeing Through Darkness: Visual Localization at Night via Weakly Supervised Learning of Domain Invariant Features. IEEE Trans Multimedia 25:1713–1726
7. Wang Meiqi, Tianqi Su, Chen Siyi et al (2023) Automatic Model-Based Dataset Generation for High-Level Vision Tasks of Autonomous Driving in Haze Weather. IEEE Trans Industr Inf 19(8):9071–9081
8. Goodfellow I J, Shlens J, Szegedy C (2015) “Explaining and Harnessing Adversarial Examples,” in Proc Int Conf Learn Represent (ICLR)
9. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P (2018) “Ensemble Adversarial Training: Attacks and Defenses,” in Proc Int Conf Learn Represent (ICLR), Sept 2018
10. Dong Y et al (2018) “Boosting Adversarial Attacks with Momentum,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Jun 2018, pp 9185-9193
11. Y Dong T, Pang H, Su, Zhu J (2019) “Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Jun 2019, pp 4312-4321
12. Esmaeili A, Edraki M, Rahnavard N, Mian A, Shah M (2024) Low-Rank and Sparse Decomposition for Low-Query Decision-Based Adversarial Attacks. IEEE Trans Inf Forensics Secur 19:1561–1575
13. Rashid A, Such JM (2023) MalProtect: Stateful Defense Against Adversarial Query Attacks in ML-Based Malware Detection. IEEE Trans Inf Forensics Secur 18:4361–4376
14. Kurakin A, Goodfellow I J, Bengio S (2017) “Adversarial Examples in the Physical World,” in Proc Int Conf Learn Represent (ICLR), Sept 2017
15. Szegedy C et al (2014) “Intriguing Properties of Neural Networks,” in Proc Int Conf Learn Represent (ICLR), Sept 2014
16. Xie C et al (2019) “Improving Transferability of Adversarial Examples With Input Diversity,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, pp 2725-2734
17. Naseer M, H Khan S, Rahman S, Porikli F (2018) “Task-generalizable Adversarial Attack Based on Perceptual Metric,” arXiv:1811.09020
18. Ganeshan A, V B S, Radhakrishnan V B (2019) “FDA: Feature Disruptive Attack,” in Proc ICCV, Feb. 2019, pp 8068–8078
19. Wang X, He K (2021) “Enhancing the Transferability of Adversarial Attacks through Variance Tuning,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Nov. 2021, pp 1924-1933
20. Wang Z, Guo H, Zhang Z, Liu W, Qin Z, Ren K (2021) “Feature Importance-aware Transferable Adversarial Attacks,” in Proc ICCV, Feb. 2021, pp 7619-7628
21. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A, “Adversarial Examples Are not Bugs, They Are Features,” in Proc Adv Neural Inf Process Syst (NIPS), Dec 2019, pp 125-136
22. Wu W et al (2020) “Boosting the Transferability of Adversarial Samples via Attention,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Aug 2020, pp 1158-1167
23. Y Qian et al (2022) “Visually Imperceptible Adversarial Patch Attacks,” Comput Secur, vol 123, Dec 2022
24. Zou J, Pan Z, Qiu J, Liu X, Rui T, Li W (2020) “Improving the Transferability of Adversarial Examples with Resized-Diverse-Inputs, Diversity-Ensemble and Region Fitting,” in Proc Eur Conf Comput Vis (ECCV), pp 563-579

25. Huang T, Menkovski V, Pei Y, Wang Y, Pechenizkiy M (2022) “Direction-aggregated Attack for Transferable Adversarial Examples,” J Emerg Technol Comput Syst, vol 18, no 3, Apr 2022
26. Lin J, Song C, He K, Wang L, Hopcroft J E (2020) “Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks,” in Proc Int Conf Learn Represent (ICLR), Sept 2020
27. Lu Y et al (2020) “Enhancing Cross-Task Black-Box Transferability of Adversarial Examples With Dispersion Reduction,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Jun 2020, pp 940-949
28. Liu Y, Chen X, Liu C, Song D (2016) “Delving into Transferable Adversarial Examples and Black-box Attacks,” in Proc Int Conf Learn Represent (ICLR), Nov 2016
29. Li Y et al (2020) Learning Transferable Adversarial Examples via Ghost Networks. Proc AAAI Conf Artif Intell 34(7):11458–11465
30. Xiong Y, et al (2022) “Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Sept 2022, pp 14963-14972
31. Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, Xingjun Ma, (2020) “Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets,” in Proc Int Conf Learn Represent (ICLR), Sept 2020, pp 5025-5034
32. Ding X, Cheng Y, Luo Y, Li Q, Gope P (2023) Consensus Adversarial Defense Method Based on Augmented Examples. IEEE Trans Ind Informatics 19(1):984–994
33. Lau CP, Liu J, Souri H, Lin W, Feizi S, Chellappa R (2024) Interpolated Joint Space Adversarial Training for Robust and Generalizable Defenses. IEEE Trans Pattern Anal Mach Intell 45(11):13054–13067
34. Yin J, Chen B, Zhu W, Chen B, Liu X (2023) Push Stricter to Decide Better: A Class-Conditional Feature Adaptive Framework for Improving Adversarial Robustness. IEEE Trans Inf Forensics Secur 18:2119–2131
35. Liu D, Wu LY, Li B, Boussaïd F, Bennamoun M, Xie X, Liang C (2024) Jacobian norm with Selective Input Gradient Regularization for interpretable adversarial defense. Pattern Recognit 145
36. Li Q, Hu Q, Lin C, Wu D, Shen C (2023) Revisiting Gradient Regularization: Inject Robust Saliency-Aware Weight Bias for Adversarial Defense. IEEE Trans Inf Forensics Secur 18:5936–5949
37. Laykaviriyakul P, Phaisangittisagul E (2023) Collaborative Defense-GAN for protecting adversarial attacks on classification system. Expert Syst Appl 214:118957
38. Zhao C, Mei S, Ni B, Yuan S, Yu Z, Jun Wang (2024) “Variational Adversarial Defense: A Bayes Perspective for Adversarial Training,” IEEE Trans Pattern Anal Mach Intell, vol 46, no 5, pp 3047-3063
39. Niu Z, Yang Y (2023) Defense Against Adversarial Attacks with Efficient Frequency-Adaptive Compression and Reconstruction. Pattern Recognit 138
40. Han K, Xia B, Li Y (2022) (AD)2: Adversarial domain adaptation to defense with adversarial perturbation removal. Pattern Recognit 122
41. Wang Y, Li X, Yang L, Ma J, Li H (2024) ADDITION: Detecting Adversarial Examples With Image-Dependent Noise Reduction. IEEE Trans Dependable Secur Comput 21(3):1139–1154
42. Nowroozi E, Mohammadi M, Golmohammadi P, Mekdad Y, Conti M, Uluagac S (2024) Resisting Deep Learning Models Against Adversarial Attack Transferability via Feature Randomization. IEEE Trans Serv Comput 17(1):18–29
43. Nowroozi E, Mohammadi M, Savas E, Mekdad Y, Conti M (2023) Employing Deep Ensemble Learning for Improving the Security of Computer Networks Against Adversarial Attacks. IEEE Trans Netw Serv Manag 20(2):2096–2105
44. Barni M, Nowroozi E, Tondi B (2020) Improving the security of image manipulation detection through one-and-a-half-class multiple classification. Multim Tools Appl 79(3–4):2383–2408
45. Kim WJ, Hong S, Yoon SE (2022) “Diverse Generative Perturbations on Attention Space for Transferable Adversarial Attacks,” in Proc IEEE Int Conf Image Process (ICIP), Oct 2022, pp 281-285
46. Hang J, et al (2022) “Ensemble adversarial black-box attacks against deep learning systems,” Pattern Recogn, vol 101, May 2022
47. Allen-Zhu Z, Li Y (2023) “Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning,” in Proc Int Conf Learn Represent (ICLR), Sept 2023
48. Xu Y, Qiu X, Zhou L, Huang X (2023) Improving BERT Fine-Tuning via Self-Ensemble and Self-Distillation. J Comput Sci Technol 38(4):853–866
49. Ma N, Zhang X, Zheng H, Sun J (2018) “ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design,” in Proc Eur Conf Comput Vis (ECCV), pp 122-138
50. Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, Jian Sun (2021) “RepVGG: Making VGG-Style ConvNets Great Again,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Dec 2021, pp 13733-13742
51. Sandler M, Howard AG, Zhu M, Zhmoginov A, Chen L (2018) “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proc IEEE Conf Comput Vis Pattern Recognit (CVPR), Dec 2018, pp 4510-4520

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Shuyan Cheng** received the B.Eng degree in computer science and technology from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2021. She is pursuing the Ph.D. degree in computer science and technology from Nanjing University of Posts and Telecommunications. Her research interests include deep learning and adversarial attacks.

**Peng Li** received the Ph.D. degree in computer science and technology from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2013. He is currently a Professor and Doctoral Supervisor with the School of Computer Science, Software and Cyberspace Security, Nanjing University of Posts and Telecommunications. He has presided over ten national, provincial and ministerial projects. His main research interests include computer communication networks, wireless sensor networks, and information security. He is a member of the CCF, the IEEE and the IEEE Communications Society.



**Jianguo Liu** received the Engineer degree from Hebei GEO University, Shijiazhuang, China, in 2021. He received the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China in 2024.

**He Xu** received the M.Eng. and Ph.D. degree in information network from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009 and 2012, respectively. He is currently a Professor and Master Supervisor with the School of Computer Science, Nanjing University of Posts and Telecommunications. His main research interests include Internet of Things (IoT) technology and applications. He is a senior member of the CCF.

**Yudong Yao** received the B.Eng. and M.Eng. degrees in electrical engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 1988. From 1987 to 1988, he was a Visiting Student with Carleton University, Ottawa, ON, Canada. From 1989 to 2000, he was with Carleton University, Spar Aerospace Ltd., Montreal, ON, Canada, and Qualcomm Inc., San Diego, CA, USA. Since 2000, he has been with the Stevens Institute of Technology, Hoboken, NJ, USA, where he is currently a Professor and the Chair with the Department of Electrical and Computer Engineering. He holds one Chinese patent and 13 U.S. patents. His research interests include wireless communications, machine learning and deep learning techniques, and healthcare and medical applications. For his contributions to wireless communications systems, he was elected as a Fellow of the National Academy of Inventors, in 2015, and the Canadian Academy of Engineering, in 2017. He has served as an Associate Editor for the IEEE Communications Letters, from 2000 to 2008, and the IEEE Transactions on Vehicular Technology, from 2001 to 2006. From 2001 to 2005, he has also served as an Editor for the IEEE Transactions on Wireless Communications.