Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17750-3

Imperceptible adversarial attack via spectral sensitivity of human visual system

Chen-Kuo Chiang[1] · Ying-Dar Lin[2] · Ren-Hung Hwang[3] · Po-Ching Lin[4] · Shih-Ya Chang[4] · Hao-Ting Li[4]

Received: 19 March 2023 / Revised: 3 November 2023 / Accepted: 25 November 2023

Abstract
Adversarial attacks reveals that deep neural networks are vulnerable to adversarial examples. Intuitively, adversarial examples with more perturbations result in a strong attack, leading to a lower recognition accuracy. However, increasing perturbations also causes visually noticeable changes in the images. In order to address the problem on how to improve the attack strength while maintaining the visual perception quality, an imperceptible adversarial attack via spectral sensitivity of the human visual system is proposed. Based on the analysis of human visual system, the proposed method allows more perturbations as attack information and re-distributes perturbations into pixels where the changes are imperceptible to human eyes. Therefore, it presents better Accuracy under Attack(AuA) than existing attack methods whereas the image quality can be maintained to the similar level as other methods. Experimental results demonstrate that our method improves the attack strength of existing adversarial attack methods by adding 3% to 23% while mostly maintaining the visual quality of SSIM lower than 0.05.

Keywords Imperceptible adversarial attack · Spectral sensitivity · Human visual system · Deep learning

1 Introduction

Deep Neural Networks (DNNs) have been thriving in computer vision and natural language processing. However, researchers have discovered that DNNs are vulnerable to adversarial examples in recent years. Adversarial examples are those specialized inputs created by the purpose of confusing a neural network, resulting in the misclassification of a given input. An Adversarial Attack is the method to find a perturbation to the input that changes the prediction of a machine learning model. The perturbation can be very small and imperceptible to human eyes. Adversarial attacks were previously proposed by Szegedy et al. [1]. A slight perturbation can be found by maximizing the network's prediction error to cause the network to misclassify.

✉ Chen-Kuo Chiang
ckchiang@cs.ccu.edu.tw

Extended author information available on the last page of the article

Published online: 20 December 2023

Recent researches [2–5] aim to create adversarial examples that cause misclassification while being imperceptible to human vision. Large perturbations enhance the attack strength but usually lead to noticeable changes in an image. Hence, adversarial attack methods have to make a trade-off between increasing attack strength and maintaining visual quality of images.

The existing methods can be roughly divided into two categories. One attempts to simulate the response of human eye to color changes through l_p-norm distance (l_0 [6], l_2 [4, 7], and l_∞ [8, 9]). However, some works [10, 11] point out that l_p-norm of the perturbations in RGB space cannot be consistent with human vision. The other is focused on controlling perturbation weight. [12, 13] add a different amount of perturbation to each pixel by adjusting the weighted value. Therefore, more perturbations can be added in the dark or complex texture area, whereas fewer perturbations are introduced in smooth area or area that attracts visual attention. [14] adjusts the perturbation weight through the just noticeable difference (JND) in the constraint of a distortion function. This improves the images quality and guarantees high image fidelity. However, these methods can be applied with just a few existing attack methods. This motivates us to propose a meta algorithm that can be easily combined with existing methods while considering human perceptual property to boost their attack accuracy. Recent work [15] enhances adversarial examples of transferability in the black-box setting via variance tuning. Gradient variance from the previous iteration is exploited to tune the gradient in current iteration to stabilize the gradient direction. It can generate visually similar adversarial images with higher transferability. Later on, [16] proposes adversarial training framework to learn attack parameters for adversarial example generation. Although these framework is flexible and can be used to boost the existing attack methods, the human perceptual property is not considered in the loop of adversarial example generation.

To address the aforementioned problems, an adversarial attack method based on the *Spectral Sensitivity* (SS) of the human visual system (HVS) is proposed. The analysis of spectral sensitivity helps to determine the less sensitive pixels where more perturbations can be introduced. This benefits the strength of adversarial attack methods. On the other hands, spectral sensitivity can also determine sensitive pixels where perturbations should be reduced. This maintains the visual quality of image when attack method is applied. In the proposed SS attack method, the color sensitivity function calculates the pixel-wise sensitivity values for the input image. Then, Bezier curve is exploited to provide smooth and continuous sensitivity estimation. Lastly, adjusted sensitivity values are added up to perturbation weight of the existing gradient based attack method. The proposed method utilizes human perceptual property, Spectral Sensitivity (SS) of the human visual system (HVS), into the attack framework. This makes it possible to relocate the perturbations into pixels where changes are imperceptible to human eyes. Therefore, the adversarial examples can be more visually consistent to humane eyes with less visual artifacts. The adversarial examples generated from our approach retain similar visual quality as the original attack methods. In addition, our method serves as a meta algorithm that can be easily combined with existing attack methods. Experimental results demonstrate that our method enhances the attack strength for white-box and the black-box attack settings on two publicly available datasets.

In summary, the contributions of this work are three folds:

1. Spectral sensitivity of HVS is proposed to improve the adversarial attack methods.
2. The proposed sensitivity-based framework is flexible can be integrated with existing gradient-based adversarial attack methods easily.
3. Experimental results demonstrate that by employing the spectral sensitivity attack, the attack strength can be enhanced while maintaining the visual quality over the existing attack methods on two benchmark datasets.

The remaining organisation is listed as follows. Section 2 overviews related works of adversarial attack methods. The proposed method is described in Section 3. Results and analysis are given in Section 4. The research is concluded in Section 5.

2 Related work

2.1 Threat model

Dong et al. [17] define the threat model by three aspects: adversary goals, adversary capabilities, and adversary knowledge. Adversary goals represent the adversarial examples used for untargeted attack or targeted attack. Untargeted attack aims to misclassify the test sample to arbitrary wrong classes, whereas targeted attack misleads the target model toward a specific wrong class. Adversary capabilities define how to generate perturbations by constrained based method or optimization based method. The adversary knowledge indicates the understanding level of the target model. There are four levels defined for adversary knowledge: white-box attack, transfer-based black-box attack, score-based black-box attack, and decision-based black-box attack. This paper aims at untargeted attacks and employs attack methods, including constrained based and optimization based attack methods. White-box and transfer-based black-box attacks are exploited in our experiments, which are usually considered in recent adversarial attack researches.

2.2 Gradient-based adversarial attack

Four gradient-based adversarial attacks are introduced in this section, including Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD).

2.2.1 Fast gradient sign method

Fast Gradient Sign Method (FGSM) [9] is the first adversarial attack method that generates an adversarial example with constrained perturbation. FGSM calculates the gradient from the loss function of the input image. The sign of the gradient is multiplied by a weighting parameter ϵ to generate perturbation noises. The perturbation is then added up to the original image as the adversarial example. The equation of the adversarial example generation is defined as

$$x' = x + \epsilon \cdot \mathrm{sign}(\nabla_x J(\theta, x, y)), \tag{1}$$

where $x\prime$ is the adversarial example, x is the input image, y is the ground-truth label, ϵ is the perturbation weight, J is the loss function of the target model, and θ is the model parameters. The gradient of loss function J is defined by ∇_x.

2.2.2 Basic iterative method

Basic Iterative Method (BIM) [8] is an extension of FGSM to perform the algoritm iteratively. BIM applies clipping after each iteration to ensure the values are constrained to the ϵ-neighbourhood of the original image. The objective function of BIM is written as

$$\begin{aligned} x'_0 &= x, \\ x'_{t+1} &= \mathrm{clip}_{x,\epsilon}\{x'_t + \alpha\, \mathrm{sign}(\nabla_x J(\theta, x'_t, y)\}, \end{aligned} \tag{2}$$

where the adversarial example of iteration $t + 1$ is obtained by adding the perturbation to x'_t with the clip operation. The perturbation is the gradient with respect to the loss function of x'_t and the ground-truth label y. The parameter ϵ in BIM is considered as a perturbation budget, which is different from that in FGSM.

2.2.3 Projected gradient descent

Projected Gradient Descent (PGD) [18] is also an iterative extension of FGSM. There are two main differences between PGD and BIM. One is that the initial adversarial example of PGD does not use the original image but the original image with random noises. In addition, PGD constraints the perturbation on l_p-ball instead of the clip operation used by BIM. The objective function of PGD is defined as

$$x'_{t+1} = \Pi_{x+S}(x'_t + \alpha \operatorname{sign}(\nabla_{x'_t} J(\theta, x'_t, y))), \tag{3}$$

where Π_{x+S} is a projection operator with perturbation set $x + S$, and α is a gradient step size. In Eq. 3, the perturbation in each iteration is projected to l_p-ball and constrained in the range of S.

Recently, Jia et al. [16] proposes adversarial training framework to learn attack parameters for adversarial example generation. Most previous methods adopt PGD with manually selected parameters for adversarial example generation. This lacks flexibility and only one attack strategy is employed. The learning framework learns to produce attack strategies and can be applied to each stage. Although the learning framework works well for high test robustness, the adversarial examples are not guaranteed and evaluated for human vision consistency.

2.3 Optimization-based method

Carlini and Wagner-l_2 attack (C&W-l_2) [7] minimize the l_2 norm of adversarial perturbations and enlarge the classification confidence gap between the incorrect classes and the ground-truth class. Therefore, C&W-l_2 is considered as a constrained optimization perturbations. The objective function is defined as

$$\min_{\omega} \|x' - x\|_2^2 + \lambda f(x'), \tag{4}$$

where the term $\|x' - x\|_2^2$ constrains perturbation by l_2-norm, and the $f(x')$ term enlarges the confidence gap. λ is a controlling weight of perturbation. $f(x')$ is defined as

$$f(x') = \max\left\{\max_{i \neq y} Z(x') - Z(x')_y, -\kappa\right\}, \tag{5}$$

where $Z(x')$ represents the value from softmax layer, which can be regarded as the confidence of the classification result. $Z(x')_y$ represents the value of correct class. The highest confidence of incorrect class is represented by $\max_{i \neq y} Z(x')$. This term $f(x')$ enforces that the distance from the highest confidence of incorrect class to the confidence of correct class should be large than the margin κ.

2.4 Perceptual color and sensitivity

l_p-norm is commonly adopted to measure the difference between two images while generating adversarial examples. However, the results measured by l_p-norm in RGB space are poorly aligned with the visual properties of human vision system. Zhao et al. [10] propose a modified $C\&W$ method based on perceptual color. It maintains adversarial strength, while producing larger RGB perturbation of imperceptibility. Experimental results show that perceptual color distance made a contribution that create adversarial images imperceptible to human eyes in smooth, saturated regions. It is also robust to two transformation-based defense methods, JPEG compression and bit-depth reduction. However, this method is directly extended from C&W. It requires further modification to enhance other types of attack method. In addition, the perturbations produced by this method does not perform well in smooth regions with low saturation. Luo et al. [13] define the perturbation sensitivity of each pixel by calculating the standard deviation of the pixel and its eight neighbors. The reciprocal of the standard deviation is obtained as the sensitivity value. Perturbation sensitivity is used to calculate the perturbation priority. Finally, a greedy algorithm is adopted to search twenty highest priority pixels in each iteration and perturb the given pixels until the total perturbation exceeds the threshold. This method is highly reliable to noises. Therefore, it is generally applicable for a large amount of applications. Some potential problems may occur in the method. The method is only verified on images of low resolution. The effectiveness of high-resolution image datasets, such as ImageNet [19], is not clear. In addition, the time complexity of greedy algorithm is high while the attack process is conducted on large-scale image datasets. Overall, these methods are not easy to adapt or combined with the existing attack methods. Recent method [20] achieves imperceptibility by limiting perturbations within high frequency components. This ensures perceptual similarity between adversarial examples and original images. Another research direction [21] is to train Invertible Neural Networks (AdvINN), which generates class-specific semantic information of the target class into the adversarial examples and dropps existing details of the original class.

3 Imperceptible adversarial attack via spectral sensitivity

In this section, an imperceptible adversarial attack method is proposed based on the spectral sensitivity of the human visual system. The proposed scheme enhances the attack strength of existing attack methods while reducing significant visible perturbation on attacked images.

3.1 Overview

In the process of adversarial attacks, the perturbation is generated by an attack method and the target model. It is usually scaled by a weighting parameter, such as $epsilon$ in Eq. 1, and then added to the original image. The adversarial example with attack information serves as an input to the target model for classification. In the conventional attack methods, the perturbation weight of each pixel is the same. In other words, all pixel values are adjusted by a single weighting parameter. However, color sensitivity is different to human vision system. For some colors, human vision can easily observe the change when a bit of perturbation is added. On the other hand, the change of some colors can be unaware if the color has low sensitivity to human vision. Based on this human vision property, spectral sensitivity is employed in our method to adjust the perturbation weight of each pixel separately. More perturbation can be added to low sensitivity pixels to enhance the attack strength, while it can be further reduced for high sensitivity pixels.

The proposed spectral sensitivity attack method is depicted in Fig. 1. In the first step, the input images are evaluated by the color sensitivity function to obtain the sensitivity matrix S.

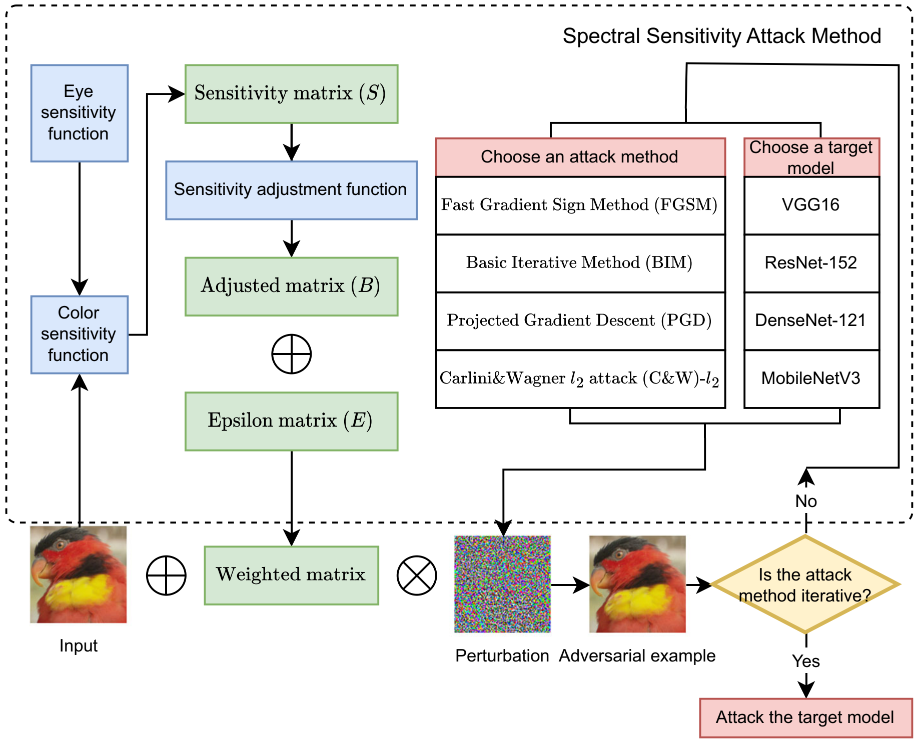


Fig. 1 Overview of System Framework. The bottom part of the figure depicts process of generating adversarial examples. Our method enclosed by the dash-line box in the upper of the figure controls the per-pixel weight of perturbation

Denote the sensitivity value s_{ij} of pixel x_{ij}. Next, the sensitivity matrix is converted through a sensitivity adjustment function as the adjusted matrix B. Afterwards, the adjusted matrix is added with the original weight of the attack method, which controls the amount of perturbation for each pixel. The adjusted matrix B aims to decrease the weights of high-sensitivity pixels to reduce the attack strength while increasing the weights of low-sensitivity pixels to enhance the attack strength. If the attack method is iterative, the above steps will repeat. Otherwise, the procedure completes. In the following sections, the details of color sensitivity function, sensitivity adjustment curve, and how to combine our scheme with existing adversarial attack methods are introduced.

3.2 Color sensitivity function

The color sensitivity function calculates the sensitivity value of each pixel by linear combination according to the Grassman's Laws [22]. Grassman's Laws describes that the response of human vision to a color light is linear, which can be obtained from a linear combination of monochromatic light, like RGB. The spectral sensitivity values of three monochromatic lights RGB are obtained from the table of the eye sensitivity function CIE1978 [23] published by the International Commission on Illumination (CIE). The three sensitivity values (Red, Green, Blue) are defined as (0.012526, 0.895494, 0.09198). In visual neuroscience, spectral sensitivity means the different reactions of photopigment in the rod cells and cone cells in the eye's retina. There is only one type of rod cell, which is mainly responsible for perceiving the luminance of light. There are three types of cone cells, which are responsible for perceiving RGB color.

The perception of human vision is affected by combining the four types of cells. This function considers the perception of four cells and is formulated to accurately describe the spectral sensitivity of the human eye. The input image can be transformed into a sensitivity matrix $S = [s_{ij}]$ via color sensitivity function defined as

$$s_{ij} = \frac{x_{ij}^{r}}{255} \times 0.012526 + \frac{x_{ij}^{g}}{255} \times 0.895494 + \frac{x_{ij}^{b}}{255} \times 0.09198, \tag{6}$$

where the RGB value of each pixel x_{ij} is divided by 255 individually and regarded as the ratio of each channel. Then, each ratio is multiplied by the spectral sensitivity value corresponding to each RGB channel. Finally, three values are added to obtain the final spectral sensitivity s_{ij} of the pixel, ranging from 0 to 1.

3.3 Sensitivity adjustment function

In the next step, sensitivity adjustment function is proposed to convert the color sensitivity value s_{ij} in Eq. 6 into the adjusted weight b_{ij}. The adjusted weights are used to be added into the original weights and aims to control the per-pixel perturbation. The adjusted weights can be negative, zero, or positive values. When negative values added to the perturbation weight, it means to weaken the attack strength and reduce the influence of the perturbation to visual perceptibility. On the other hand, positive values increases the amount of perturbation and thus enhances the attack strength.

Three color sensitivity ranges are defined for adjusted weights: *high-sensitivity*, *medium sensitivity* and *low-sensitivity*. Assume that the sensitivity values are uniformly distributed in [0, 1], s_{ij} greater than 0.7 can be defined as a high sensitivity pixel. For high sensitivity pixels, a little perturbations are easily perceptible by human vision system. Thus, a negative adjusted weight b_{ij} should be returned by the sensitivity adjustment function. This reduces the amount of perturbation that is allowed as attack information. On the contrary, s_{ij} less than 0.4 is regarded as a low sensitivity pixel. Since the pixel is less sensitive to human vision, more attack information can be added. Sensitivity adjustment function returns a positive weight to increase the amount of attack information. Otherwise, the pixel is considered as medium sensitivity, and b_{ij} is set to 0. Notice that the range of low-sensitivity (0-0.4) is a bit larger than the range of high-sensitivity (0.7-1). This is to ensure that more perturbations can be added into the pixels with low-sensitivity. The range of attack strength enhancement is larger than that of attack strength reduction. In a way, this increases the overall attack strength. Based on the discussion above, the sensitivity adjustment function can be formulated by

$$b_{ij} = \begin{cases} -\alpha_1 s_{ij}, & \text{if } s_{ij} > 0.7 \\ 0, & 0.4 \leq s_{ij} \leq 0.7 \\ \alpha_2(1 - s_{ij}), & s_{ij} < 0.4, \end{cases} \tag{7}$$

where two parameters α_1 and α_2 are used to control the magnitude of the adjusted weights b_{ij} which affect the amount of perturbation as attack information. A total cost function $C(\alpha_1, \alpha_2)$ is defined to select the best parameters α_1 and α_2. Visual structural similarity SSIM and Accuracy under Attack (AuA) are considered in the cost function to balance the attack performance and the image quality. Both terms are ranging from 0 to 1. The cost function aims to select parameters α_1 and α_2 that performs adversarial attack to reduce the recognition accuracy while maintaining the visual quality of images as similar as possible. The overall cost function is defined as

$$C(\alpha_1, \alpha_2) = \text{AuA} + (1 - \text{SSIM}). \tag{8}$$

When minimizing the cost function, it enforces lower accuracy after attack and higher visual quality at the same time. For the definition of AuA and SSIM, please refer to Sections 4.1.1 and 4.1.3 for more details. Parameters α_1 and α_2 can be determined empirically through experiments. They are turning parameters and will be different for dataset types and image resolutions. From the experimental results, the parameter α_1 is empirically set to 0.0175 and α_2 is set to 1.2. Detailed experimental settings are provided in Table 2 of Section 4.2.

Figure 2 illustrates the results of the sensitivity adjustment function. The range of high-sensitivity, meidum-sensitivity and low-sensitivity are represented by three line segments. Note that for the range of high-sensitivity (0.7-1), line segment still has small slope and very close to zero due to a small parameter α_1. Since the sensitivity adjustment function is discontinuous at 0.4 and 0.7, the Bezier curve is adopted to smooth the function. 14 points are sampled on these three segments to calculate the Bezier curve.

The formula of sensitivity adjustment curve can be estimated by the following equation

$$P(s_{ij}) = \sum_{k=0}^{13} P_k \binom{13}{k} s_{ij}^k (1 - s_{ij})^{13-k}, \tag{9}$$

where P_k represents the points on the three line segments. Based on Eq. 9, it is straightforward to obtain b_{ij} from $P(s_{ij})$. In Fig. 2, the *Sensitivity Adjustment Curve* represents the continuously smoothing function for the original sensitivity adjustment function. This makes the range of medium-sensitivity a monotonic decreasing function and more flexible rather than using a fixed value in the original function. Mean Absolute Error (MAE) is calculated for 20 points evenly sampled from the function range 0 to 1 to compare the difference of the approximated function to the original one. The MAE is only 0.0251. In fact, the approx-

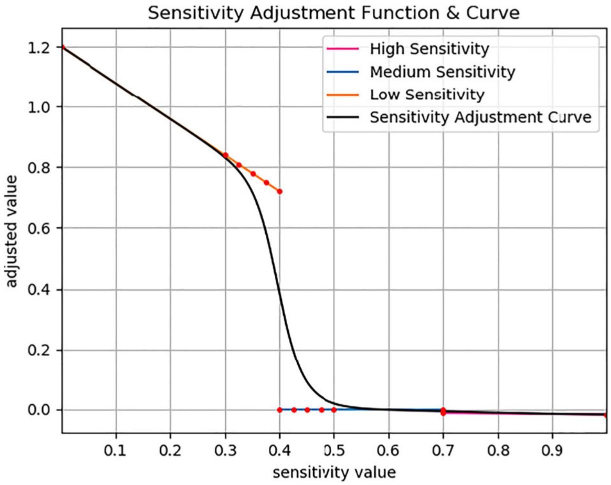


Fig. 2 Sensitivity adjustment function and curve

imation function only produces slightly different values within the range from 0.31 to 0.53 compared to the original function.

The estimated sensitivity s_{ij} on sample images are depicted in Fig. 3a. The green color represents the high-sensitivity area, blue represents the medium-sensitivity area, and red represents the low-sensitivity pixels. Figure 3a represents images of rich texture and dark background. One can notice that there are more red pixels on the feather of the peacock and leopard skin where the color changes are difficult to detect by human vision. In fact, almost all pixels are considered as medium or low sensitivity by our method in the first two images. In addition, our method can distinguish bright and dark areas. For the background in a car, they are mostly estimated as low sensitivity, while the background of a ship in Fig. 3b is almost bright colors. Such background is considered as high-sensitivity area by our method. In the rest two images in Fig. 3b, green color of the frog are more sensitive to human vision than the red cars. The pixels of frog and car are mostly estimated as medium- and low-sensitivity pixels, respectively.

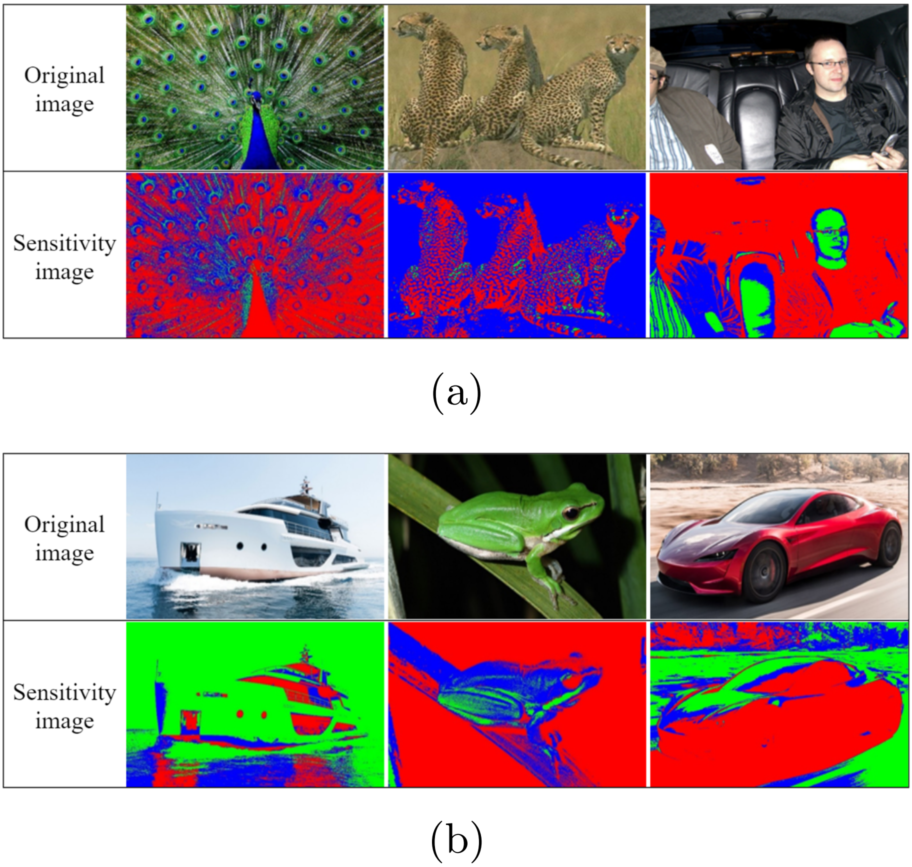


Fig. 3 Sensitivity distribution in different conditions. (a) Complicated texture and dark background, and (b) Bright background and pure color objects

3.4 Enhancing existing attack methods

Existing attack methods can be enhanced by the proposed framework to achieve imperceptible adversarial attack to human vision system base on per-pixel weight adjustment. Conventional attack methods mainly include the following components: the input x, the adversarial example x', the perturbation δ, perturbation weight ϵ and iteration index k, if it is a iterative method.

Algorithm 1 Spectral Sensitivity Adversarial Attack

```
Input:

$x$: original image, $y$: ground-truth label, $K$: number of iterations
$A$: attack method, $\epsilon$: perturbation weight
Output:
$x'$ : adversarial example
Initialize $x'_0 \leftarrow x$
for $k \leftarrow 1$ to $K$ do
    Calculate sensitivity matrix $S$ by Eq. 6
    Calculate adjusted matrix $B$ by Eq. 9
    Calculate perturbation $\delta_k \leftarrow A(x'_{k-1}, y)$
    Create epsilon matrix $E$ by $\epsilon$ by the same size as $x$
    $x'_k \leftarrow x'_{k-1} + \delta_k \bigodot (B + E)$
end for
return $x' \leftarrow x'_k$
```

The per-pixel weight adjustment scheme can be combined with existing attack methods by the following steps. Firstly, Eq. 6 is adopted to calculate the sensitivity matrix $S = [s_{ij}]$ from the RGB values of the original image. Matrix S serves as the input of Eq. 9 and is converted into adjusted matrix $B = [b_{ij}]$. The selected attack method A generate the perturbation δ_k according to x'_{k-1} and output y in iteration k. The epsilon matrix E has the same size as x is created. All entries are set to ϵ. The final weight matrix can be calculated by adding up matrix B with matrix E. Lastly, element-wise multiplication is performed by the final weight matrix and the perturbation to determine the final perturbation. New adversarial example x'_k is generated by adding x'_{k-1} and the final perturbation. The algorithm is presented in Algorithm 1.

4 Experimental results

We demonstrate the experimental results in this section. The evaluation metrics for our task are introduced in Section 4.1. The experiment settings are described in Section 4.2. The quantitative and visualization results are condicted on two evaluation datasets, CIFAR10 and ILSVRC2012, in Section 4.3 to Section 4.6, respectively.

4.1 Evaluation metrics

The section introduces three evaluation metrics, including Accuracy under Attack, Perturbation MAE and Structural Similarity Index Measure.

4.1.1 Accuracy under attack

To compare the attack strength of the original attack method and the spectral sensitivity adversarial attacks, adversarial examples are used as model input and observe the model accuracy under attacks. The equation of Accuracy under Attack (AuA) is defined as

$$\mathrm{AuA}(C, A_\epsilon) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{(\mathrm{C}(A_\epsilon(x_i))==y_i)}, \tag{10}$$

where C denotes classifier or target model, A_ϵ denotes attack method with parameter ϵ, and N denotes number of testing samples. The indicator function $\mathbb{1}$ returns one if the condition is true. More decreasing of AuA means stronger attack strength. We conduct this experiment under two kinds of adversarial settings, including white-box attacks and black-box attacks. In the white-box attacks, adversarial examples are generated by the target model. For black-box attack, ResNet152 is adopted as a substitute model to generate adversarial samples to attack three target models, DenseNet121, VGG16, and MobileNetV3.

4.1.2 Perturbation MAE

The Mean Absolute Error (MAE) of perturbations measures the difference when attack information is added into each pixel. Larger MAE means more attack information included and may introduce high visual changes in the image. The Perturbation MAE can be calculated as

$$\mathrm{MAE} = \frac{1}{MN}\sum_{i=1}^{N} \|x_i' - x_i\|_1, \tag{11}$$

where M denotes the number of total pixels in each image, N denotes the number of testing samples.

4.1.3 Structural Similarity Index Measure (SSIM)

The Structural Similarity Index Measure (SSIM) considers brightness, contrast, and structural similarity. It is used as an evaluation criterion for measuring the similarity between adversarial example and the original image. The range of SSIM is [0, 1], where 1 means the same as the original image, and 0 is entirely different from the original image.

4.2 Experimental settings

The experiments were conducted on image datasets to validate our method, including CIFAR-10 [24] and ILSVRC2012 [25]. CIFAR-10 is an established computer-vision dataset for digit recognition. It consists of 60,000 32x32 color images of ten classes. Each class has 6000 images. ILSVRC2012 is the abbreviation for ImageNet Large Scale Visual Recognition Challenge 2012. There are 1000 object categories as those in ImageNet. The training data is the subset of ImageNet, containing the 1000 categories and 1.2 million images. The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. On these two datasets, adversarial examples were crafted

Table 1 Original accuracy of target models on CIFAR-10 and ILSVRC2012

Model	CIFAR-10	ILSVRC2012
VGG16	91.00%	72.10%
ResNet152	85.34%	78.30%
DenseNet121	89.90%	74.20%
MobileNetV3	80.14%	74.24%

by eight adversarial attack methods, including FGSM, BIM, PGD, C&W-l_2, and the Spectral Sensitivity (-SS) version of these four methods. 3000 images were randomly selected from each of the two datasets for evaluation. The perturbation weight $epsilon$ is set from 0.03 to 0.3 on CIFAR-10 and 0.03 to 0.39 on ILSVRC2012 to compare the results of FGSM and SS-FGSM. For the remaining six iterative methods, the results were observed by iteration from 1 to 10. The experimental results are validated by four target models, including DenseNet121 [26], ResNet152 [27], VGG16 [28], and MobileNet-V3 [29]. The original accuracy before attack methods applied is presented in Table 1.

In order to select the best combination of parameters α_1 and α_2, 500 images of CIFAR-10 were used to generate adversarial examples. SS-FGSM is employed as the attack method and DenseNet121 is used as the target model. In Table 2, the minimum total cost 0.696698 was obtained when α_1 is set to 0.0175 and α_2 is set to 1.2. The accuracy rate using this set of parameters is 24.9%, which is lower than the 27.9% of the original FGSM. The SSIM is 0.5523, which is about 0.016 lower than that of the original FGSM. Such parameter set meets our needs to increase the attack strength yet with small visual quality degradation.

4.3 Evaluation of white-box attack

In this subsection, we present the results of the white-box attack on CIFAR-10 and ILSVRC datasets. For results on CIFAR-10 dataset, each figure shows the accuracy of four models by original attack methods and the corresponding SS-enhanced version. The x-axis of the diagram is the perturbation weight (epsilon) in Fig. 4a or the number of iterations in Fig. 4b, c, d, and the y-axis depicts the accuracy. The solid curves and the dashed curves are the results of the original attack methods and proposed SS-enhanced methods, respectively. From Fig. 4a, b, c, d, the dash-curves are all under solid curves. The demonstrates that the proposed Spectral Sensitivity method can further improve the existing attack methods on different target model. The improvement is significant, especially for FGSM, BIM and PGD methods. One can note that the accuracy tends to converge as epsilon or iteration number becomes larger. This is

Table 2 The influence of α_1 and α_2 on total cost

Total cost	$\alpha_1 = 0.0150$	$\alpha_1 = 0.0175$	$\alpha_1 = 0.0200$	$\alpha_1 = 0.0225$	$\alpha_1 = 0.0250$
$\alpha_2 = 0.9$	0.703880	0.703757	0.703947	0.704192	0.705160
$\alpha_2 = 1.0$	0.698737	0.697646	0.697837	0.699049	0.700340
$\alpha_2 = 1.1$	0.698349	0.697274	0.697789	0.698679	0.699969
$\alpha_2 = 1.2$	0.696745	0.696698	0.697535	0.698103	0.698748
$\alpha_2 = 1.3$	0.697044	0.697064	0.697258	0.697503	0.698471
$\alpha_2 = 1.4$	0.704331	0.703905	0.703778	0.704024	0.704669

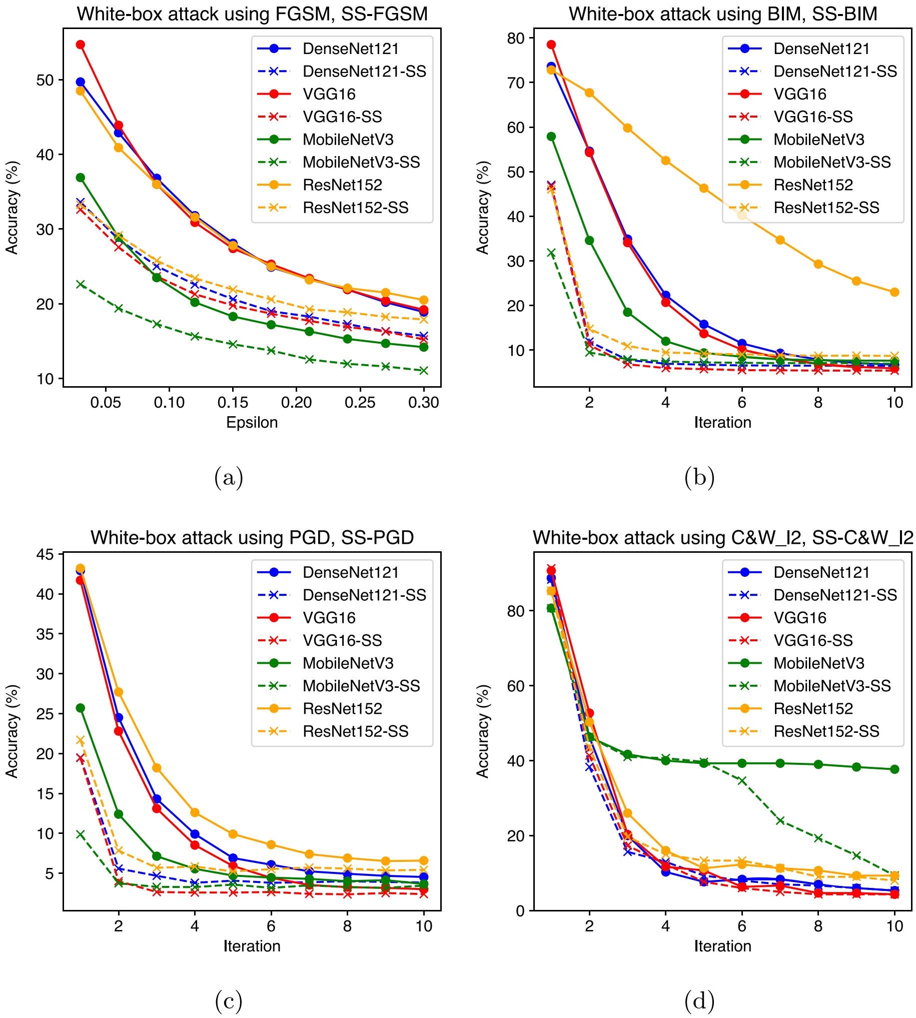
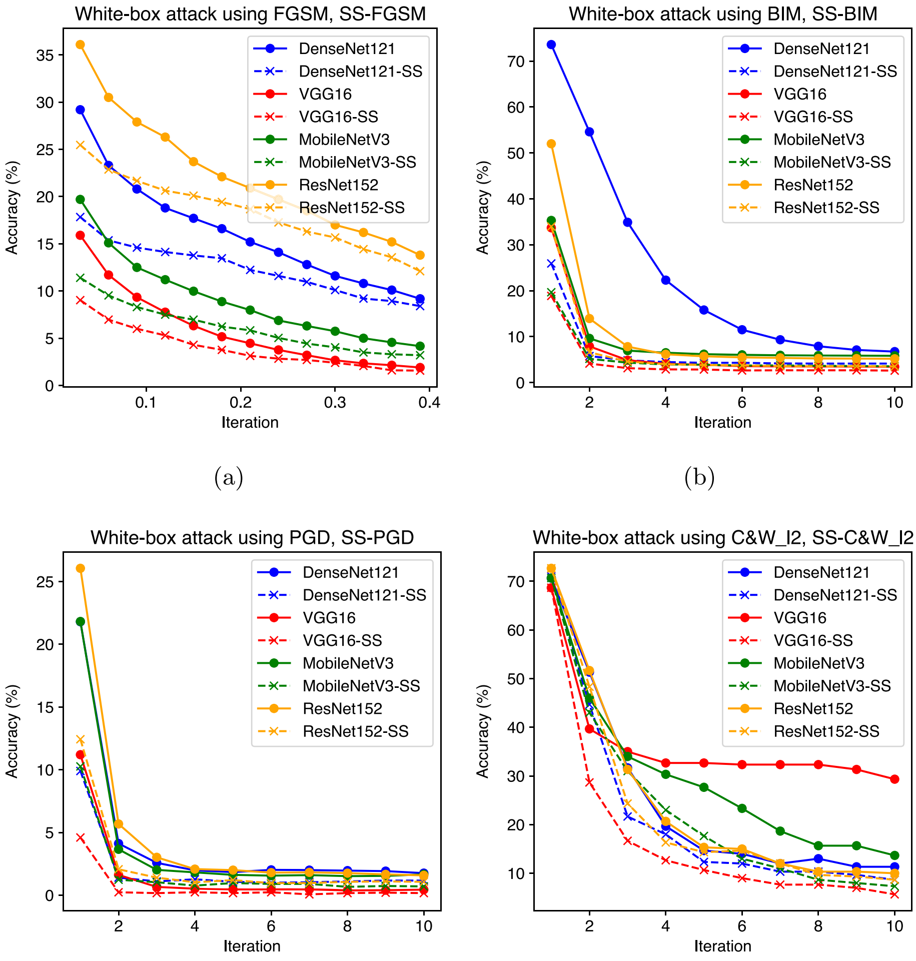


Fig. 4 Accuracy of white-box attack with different attack methods on CIFAR-10: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, (c) PGD/SS-PGD, (d) C&W/SS-C&W

because the amount of attack information also increases as the iteration number gets higher. In addition, Fig. 4d depicts that MobileNetV3 has lower accuracy degradation around 40% under the attack of C&W-l_2. When applying our sensitivity based attack method, the accuracy decreases under 20%.

Following the same settings for experiments on ILSVRC2012 dataset, four target models were attacked by original attack methods and the corresponding SS-based version. From Fig. 5 a, b, c and d, all dash curves are under the corresponding solid curve. This indicates all attack methods can be further enhanced by the proposed spectral sensitivity adversarial attack. One can notice that the accuracy on ILSVRC2012 dataset decreases significantly than CIFAR-10 in the first three iterations. Among four target models, the accuracy of ResNet152 decreases less than other models. On the contrary, VGG16 is pretty vulnerable under different attack methods.



(a)

(b)

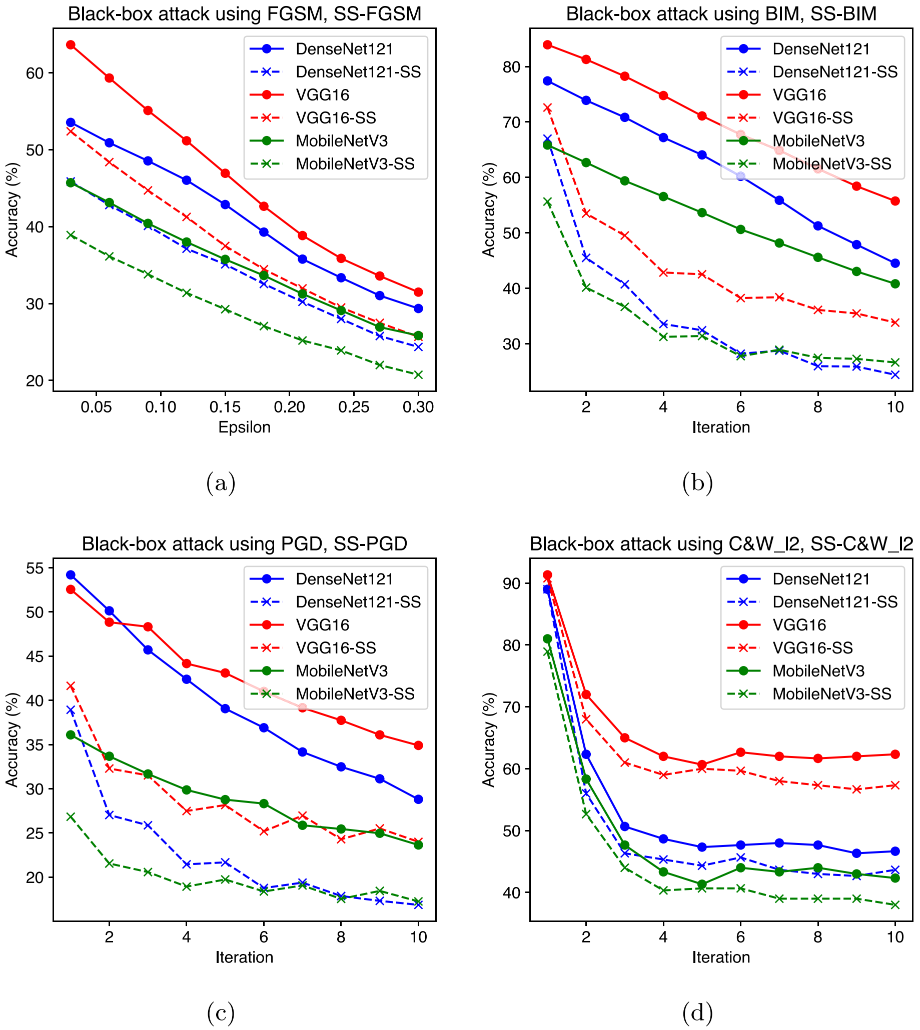
(c)

(d)

Fig. 5 Accuracy of white-box attack with different attack methods on ILSVRC2012: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, (c) PGD/SS-PGD, (d) C&W/SS-C&W

4.4 Evaluation of black-box attack

In the black-box attack experiments, ResNet152 was used to generate adversarial examples to attack other three target models: DenseNet121, VGG16, and MobileNetV3. The results of the black-box attack on CIFAR-10 presented in Fig. 6 indicate that the adversarial examples generated by the ResNet152 can also attack other models effectively. All the dash-curves of spectral sensitivity enhanced attack methods achieve lower accuracy than that of the original attack methods. According to the results, our method is able to provide more effective perturbations as attack information based on the analysis of spectral sensitivity. Among four attack methods, the dash-curve follows solid curve closely for C&W$_{l2}$. In addition, the lowest



(a)

(b)

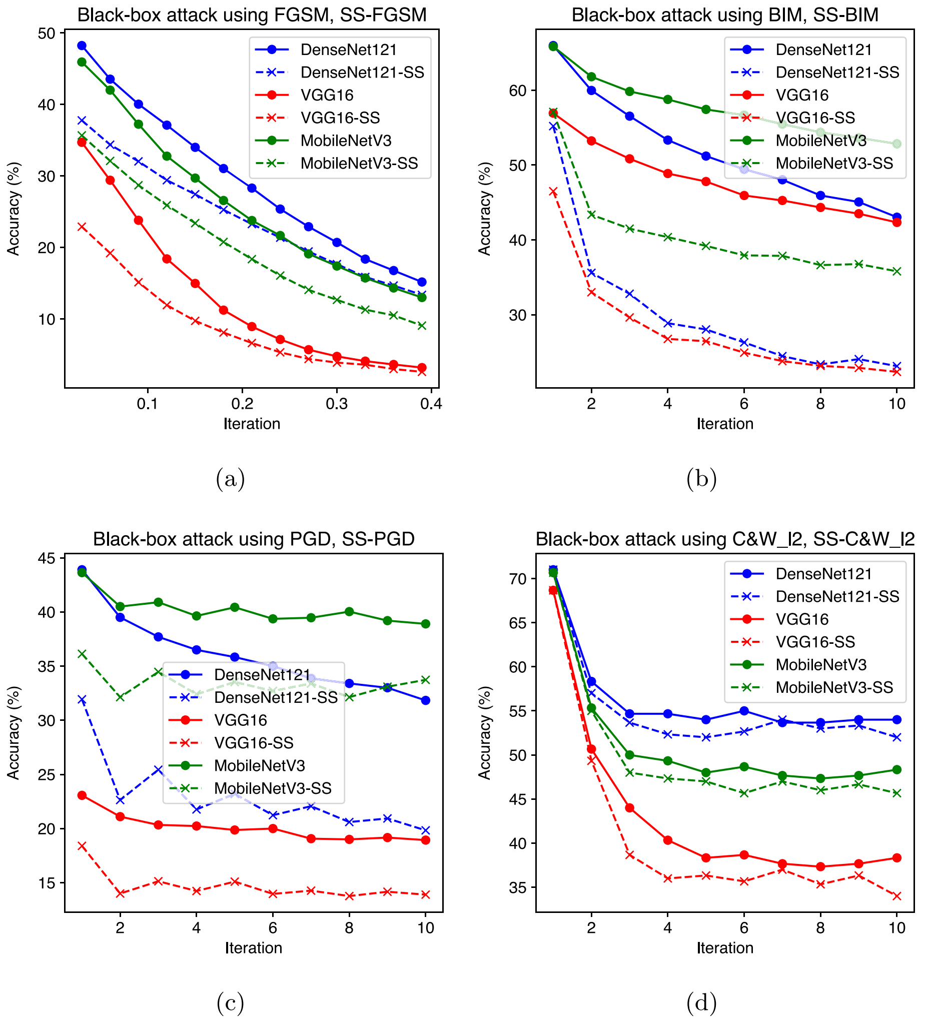
(c)

(d)

Fig. 6 Accuracy of black-box attack with different attack methods on CIFAR-10: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, (c) PGD/SS-PGD, (d) C&W/SS-C&W

accuracy under attack of C&W$_{l2}$ is higher than the other three attack methods. This shows that C&W$_{l2}$ presents relatively poor attack performance than other methods.

In Fig. 7, the black-box attack results on ILSVRC2012 also verify that the proposed spectral sensitivity based method can significantly improve the transferability of adversarial examples on large-scale image dataset over the existing attack methods. Especially for DenseNet121, the gap between dash-curve and solid-curve is wider when BIM and PGD methods are applied. This indicates our method improves BIM and PGD more than the other two methods. The possible reason is when the attack method is performed iteratively, our method has more chances to adjust attack information in each iteration.



(a)

(b)

(c)

(d)

Fig. 7 Accuracy of black-box attack with different attack methods on ILSVRC2012: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, (c) PGD/SS-PGD, (d) C&W/SS-C&W

4.5 Image quality assessment

In this experiment, adversarial examples generated by existing attack methods are compared to those generated by spectral sensitivity based methods. Perturbation MAE is used to compare the amount of attack information added to the images. In addition, SSIM is calculated to assess image quality.

Adversarial examples are generated by exploiting ResNet152 as the target model. In Fig. 8, perturbation MAE are depicted for each attack method on CIFAR-10 dataset. The amount of attack information of FGSM is controlled by the parameter ϵ whereas The amount of attack information of other methods are increased by each iteration. Thus, two separate figures are provided for FGSM and other methods in this experiments. We can notice that all dash curves

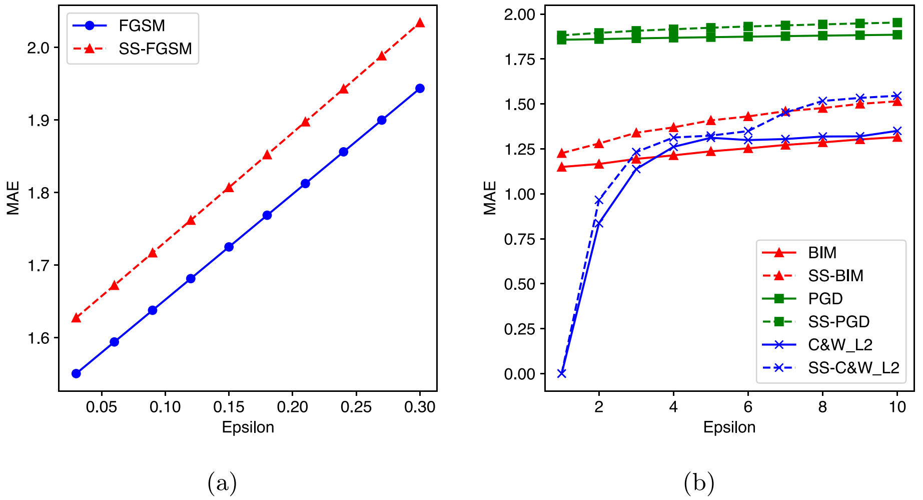


Fig. 8 Results of Perturbation MAE on CIFAR-10: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, PGD/SS-PGD, and C&W/SS-C&W

are above the solid curves by each attack method. This demonstrates that our method is able to increase the perturbation amount. On the contrary, SSIM serves as an evaluation of image quality assesment. It is adopted to calculate the similarity between adversarial examples and original images. From the results of SSIM presented in Fig. 9, the dash curves are only slightly below the solid curves. This indicates that the image quality degradation is small where more attack information can be added to the images by our method. Specifically, the maximum difference between two curves does not exceed 0.05 in each attack method. According to [30], if the SSIM is lower than 0.05, the difference between the two images is not aware by human vision.

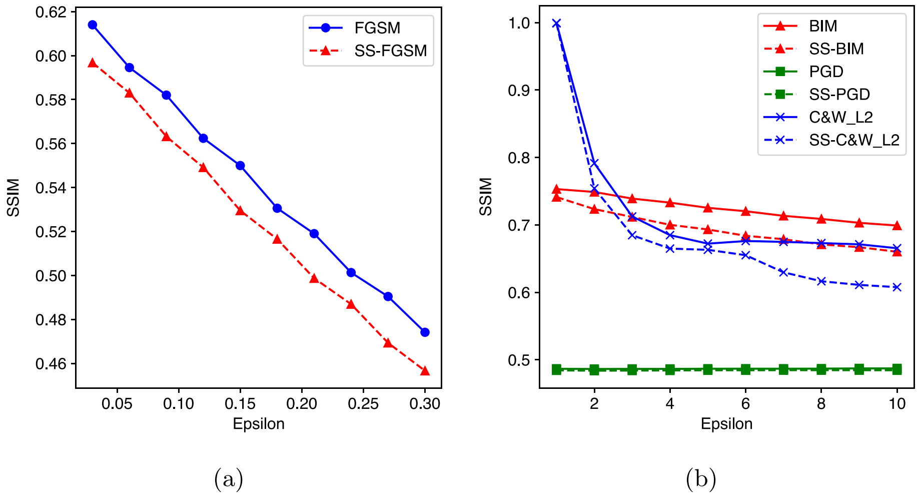


Fig. 9 Results of SSIM on CIFAR-10: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, PGD/SS-PGD, and C&W/SS-C&W

We further provide the comparison of sample visualization. The adversarial examples of CIFAR-10 generated via ResNet152 are compared by original attack method and the proposed method. In Fig. 10b, the perturbation of SS-FGSM and FGSM are visually close to each other. In Fig. 10c, the SS-BIM has the similar black perturbation spots around the fuel tank cap as BIM. However, when the iteration is greater than 8, SS-BIM is able to distribute the perturbations elsewhere, making the spot to be more imperceptible. In Fig. 10d, e, the image perturbation of SS-PGD and SS-C&W-l_2 are very similar to the original methods from the perspective of human vision. The experimental investigation on CIFAR-10 demonstrates that our method can enhance the attack strength and improve the imperceptibility of perturbation.

The amount of perturbation of the original attack method is also compared to that of spectral sensitivity based method on ILSVRC dataset. In Fig. 11, the dash curves representing the SS-based method have higher perturbation MAE in the adversarial samples on ILSVRC2012, where solid curves representing the original methods. It is very effective that SS-based methods significantly improve the attack strength over the original attack methods. In Fig. 12b, the SSIM of SS-PGD and SS-C&W-l_2 are very close to the original method. It means that the visual perception of the perturbation are very close when SS-based methods are applied. Among all attack methods, the image quality degradation of BIM is relatively larger than other methods. The visualization results on ILSVRC2012 are presented in Fig. 13. From sample images of four existing methods and corresponding SS-based method, the visual difference is hard to tell in SS-C&W-l_2 and C&W-l_2. For this sample image, perturbations primarily appear in the red area when epsilon is over 0.21 of SS-FGSM. Overall, our method is able to enhance the existing attack methods to achieve lower recognition accuracy while maintain the image quality to the similar level. Based on the analysis of spectral sensitivity, the additional amount of perturbations can be reasonably re-distributed in the image and reduces the visually impact as much as possible.

4.6 Comparison of average performance

The average of each evaluation metric is calculated to compare quantitative improvement. The results of ResNet152 were used to calculate the average white-box AuA, SSIM, and Perturbation MAE. In the black-box attack, the average AuA of DenseNet121 is presented, which is under the attack of ResNet152 adversarial examples. Results on CIFAR-10 are presented in Table 3. The proposed SS-based method increases 9.1%, 20.9%, 5.9%, and 5.5% perturbations than the original methods FGSM, BIM, PGD and C&W-l_2, respectively. Our method can also decrease the model accuracy under the knowledge of the white-box and the black-box attacks. To preserve the visual quality, the gap between our method and the original methods is less than 0.03 of SSIM, maintaining the imperceptibility level as the original attacks.

The average of four evaluation metrics are also conducted by ResNet152 for quantitative comparison on ILSVRC2012 dataset. For the black-box attack, the average accuracy of DenseNet121 is chosen, which is attacked by adversarial examples generated by ResNet152. As presented in Table 4, the SS-based methods produce 9.2%, 23.2%, 4.7% and 3.3% perturbations in MAE than the original methods FGSM, BIM, PGD and C&W-l_2, respectively. The range of introduced perturbations is wide. The possible reason is that our method manipulates perturbations directly and more suitable for gradient-based method FGSM and BIM. For constrained or optimization method, such as PGD and C&W-l_2, the perturbations are affected

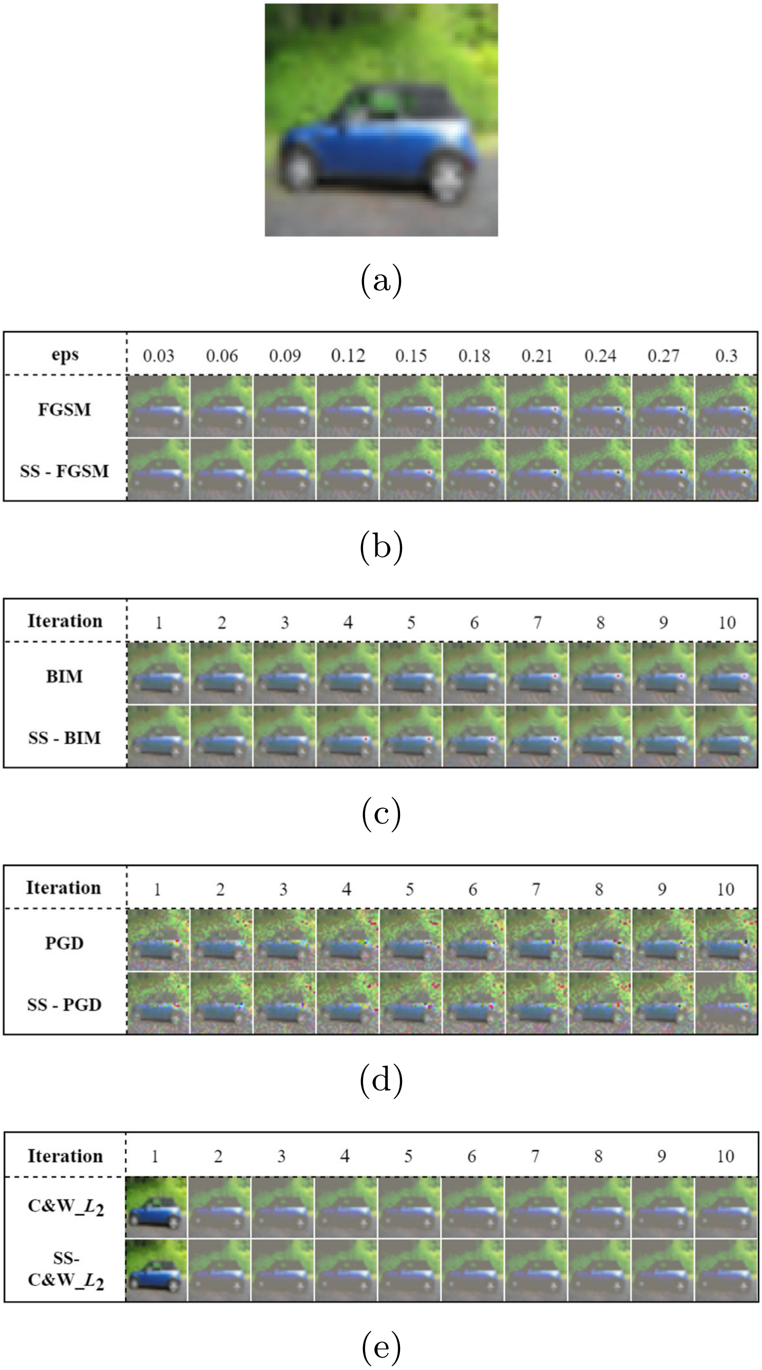


Fig. 10 Visualization of sample results on CIFAR-10: (a) the original image, and adversarial images generated by (b) FGSM/SS-FGSM, (c) BIM/SS-BIM, (d) PGD/SS-PGD, (e) C&W/SS-C&W

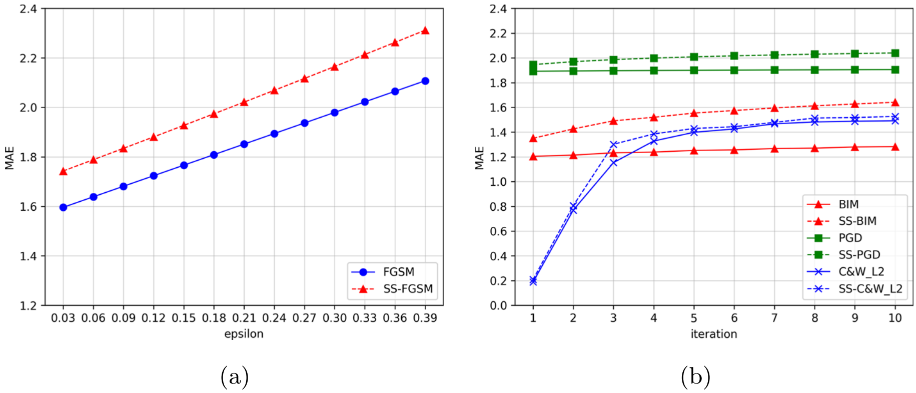


Fig. 11 Results of perturbation MAE on ILSVRC2012: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, PGD/SS-PGD, and C&W/SS-C&W

by other constraints. Therefore, the adjustment of perturbations via our approach is not that significant. SS-based methods also introduce more attack strength to achieve lower accuracy than the original methods on ILSVRC2012. On the other hand, there is minor difference of SSIM between the SS-based methods and the original attack methods.

5 Conclusions and future work

In this paper, we present an imperceptible adversarial attack method via spectral sensitivity of the human visual system. Based on the color sensitivity function, the sensitivity adjustment function is proposed to calculate the weighting parameter to adjust the initial perturbation weights. This makes it possible to increase the amount of perturbations where the changes are imperceptible to human vision system. On the other hand, our method reduces the amount

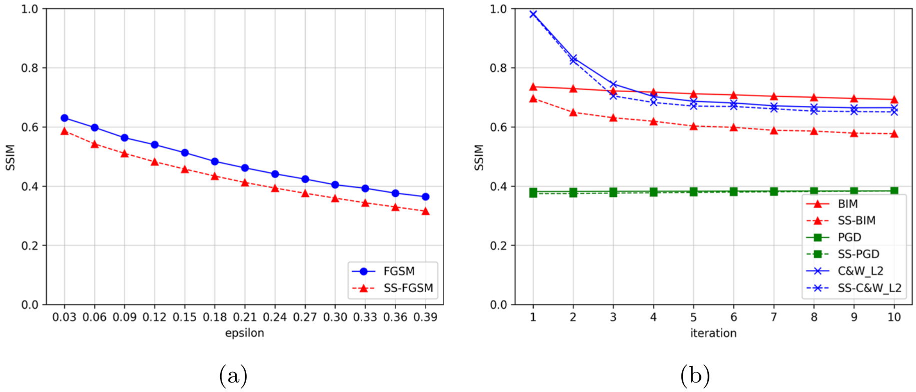


Fig. 12 Results of SSIM on ILSVRC2012: (a) FGSM/SS-FGSM, (b) BIM/SS-BIM, PGD/SS-PGD, and C&W/SS-C&W

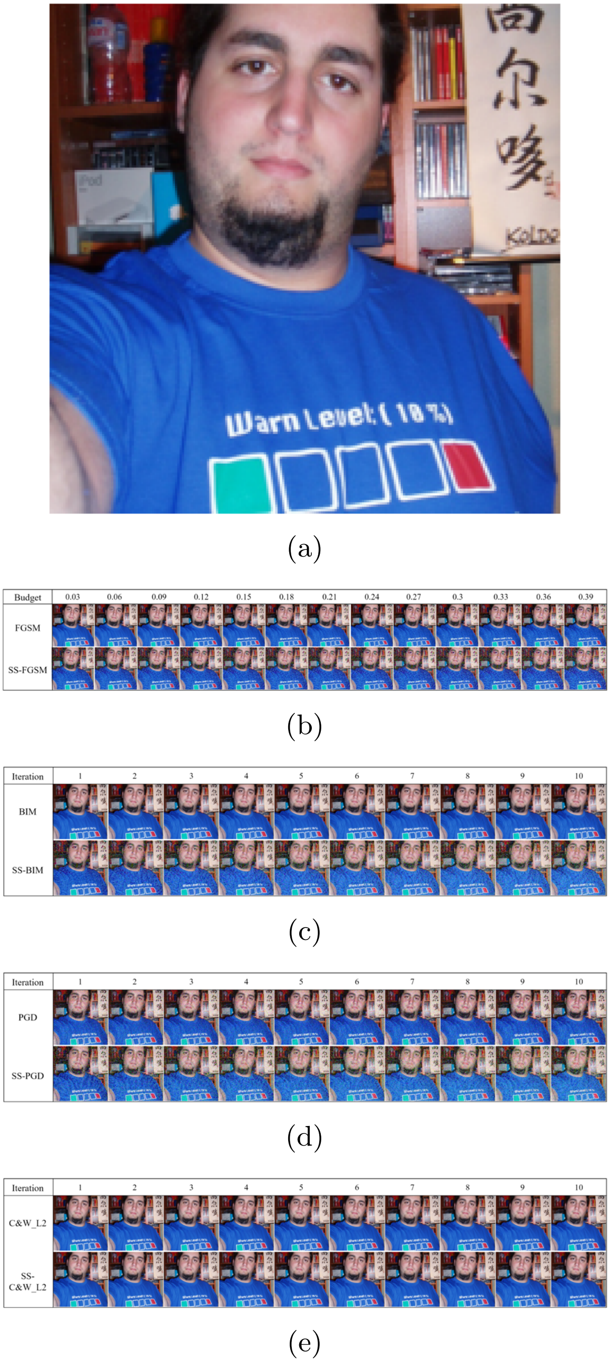


Fig. 13 Visualization results on ILSVRC2012: (a) the original image, and adversarial images generated by (b) FGSM/SS-FGSM, (c) BIM/SS-BIM, (d) PGD/SS-PGD, (e) C&W/SS-C&W

Table 3 Comparison of average AuA, SSIM and MAE on CIFAR-10

Attack methods	White-box AuA	Black-box AuA	SSIM	MAE
FGSM	29.7%	41.1%	0.51	1.75
SS-FGSM	22.8%	34.2%	0.49	1.91
BIM	45.2%	61.3%	0.64	1.24
SS-BIM	13.5%	35.2%	0.61	1.50
PGD	14.8%	39.5%	0.39	1.87
SS-PGD	7.39%	22.5%	0.38	1.98
C&W-l_2	24.2%	53.4%	0.67	1.28
SS-C&W-l_2	22.8%	49.9%	0.65	1.35

of perturbations in the sensitive area. This helps to re-distribute the perturbations in the images as the amount of attack information increases while maintain the visual quality to a similar level. Four recent attack methods and four deep learning models are included in the experiments on two image datasets. Experimental results demonstrate that our method is able to introduce more perturbations as attack information to achieve lower recognition accuracy of white-box and black-box attacks. Results of SSIM and visualization of sample results presents similar image quality to the existing attack methods. In addition, Our method can be combined with gradient-based or optimization based adversarial attack methods. From the experimental results, our method is especially beneficial to BIM and PDG which learns the attack information iteratively. It is also interesting to see that ResNet152 is more defensive than the other deep learning models in some experiments.

The data that support the findings of this study are available in CIFAR-10 [24] and LSVRC2012 [25] which are publicly available datasets. In the future, saliency detection method [31] can be combined to determine sensitive regions. Related research directions, such as camouflaged detection [32], can be also beneficial to exploit visual perception knowledge. To further improve the sensitivity scores of small regions, transformer networks designed for small object detection [33] can be employed as the backbone models. Overall, we hope to develop more general spectral sensitivity based attack methods in a learning fashion. This will lead to effect attack method on more diverse datasets and different applications. In addi-

Table 4 Comparison of average AuA, SSIM and MAE on ILSVRC2012

Attack methods	White-box AuA	Black-box AuA	SSIM	MAE
FGSM	22.2%	29.4%	0.48	1.85
SS-FGSM	18.3%	24.0%	0.43	2.02
BIM	11.2%	51.9%	0.71	1.25
SS-BIM	7.10%	30.2%	0.61	1.54
PGD	4.75%	36.1%	0.38	1.92
SS-PGD	2.33%	23.0%	0.37	2.01
C&W_l_2	24.9%	56.3%	0.73	1.22
SS-C&W_l_2	23.0%	55.1%	0.72	1.26

tion, our work is possible to bring new defense approaches for deep learning models when attack information becomes more and more imperceptible.

Acknowledgements This work was partially supported by the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and Center for Innovative Research on Aging Society (CIRAS). This work was also supported in part by the National Science and Technology Council, Taiwan under Grant NSTC 111-2634-F-006-012 and NSTC 111-2634-F-194-003.

Declarations

Conflict of interest Authors declare wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

1. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R (2014) Intriguing properties of neural networks. In: Bengio Y, LeCun Y (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings
2. Moosavi–Dezfooli S–M, Fawzi A, Fawzi O, Frossard P (2017) Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1765–1773
3. Chen J, Jordan MI, Wainwright MJ (2020) Hopskipjumpattack: A query–efficient decision-based attack. In: 2020 Ieee Symposium on Security and Privacy (sp), pp 1277–1294. IEEE
4. Moosavi–Dezfooli S–M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2574–2582
5. Papernot N, McDaniel P, Goodfellow I (2016) Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277
6. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp 372–387. IEEE
7. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: 2017 Ieee Symposium on Security and Privacy (sp), pp 39–57. IEEE
8. Kurakin A, Goodfellow IJ, Bengio S (2017) Adversarial examples in the physical world. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Workshop Track Proceedings
9. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: Bengio Y, LeCun Y (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings
10. Zhao Z, Liu Z, Larson M (2020) Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1039–1048
11. Kang D, Sun Y, Hendrycks D, Brown T, Steinhardt J (2019) Testing robustness against unforeseen adversaries. arXiv preprint arXiv:1908.08016
12. Croce F, Hein M (2019) Sparse and imperceivable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 4724–4732
13. Luo B, Liu Y, Wei L, Xu Q (2018) Towards imperceptible and robust adversarial example attacks against neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32
14. Zhang Z, Qiao K, Jiang L, Wang L, Yan B (2020) Advjnd: Generating adversarial examples with just noticeable difference. In: Machine Learning for Cyber Security, pp 463–478. Springer, ???
15. Wang X, He K (2021) Enhancing the transferability of adversarial attacks through variance tuning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19–25, 2021, pp 1924–1933. Computer Vision Foundation / IEEE, ???

16. Jia X, Zhang Y, Wu B, Ma K, Wang J, Cao X (2022) LAS-AT: adversarial training with learnable attack strategy. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022, pp 13388–13398. IEEE, ???
17. Dong Y, Fu Q–A, Yang X, Pang T, Su H, Xiao Z, Zhu J (2020) Benchmarking adversarial robustness on image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 321–331
18. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings
19. Deng J, Dong W, Socher R, Li L–J, Li K, Fei–Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255. Ieee
20. Luo C, Lin Q, Xie W, Wu B, Xie J, Shen L (2022) Frequency-driven imperceptible adversarial attack on semantic similarity. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pp 15294–15303. IEEE, ???
21. Chen Z, Wang Z, Huang J, Zhao W, Liu X, Guan D (2023) Imperceptible adversarial attack via invertible neural networks. In: Williams B, Chen Y, Neville J (eds.) Thirty–Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, pp 414–424. AAAI Press, ???
22. Grassmann H (1853) Zur theorie der farbenmischung. In: Ann Phys, pp 69–84. Wiley, ???
23. Wyszecki G, Stiles WS (1982) Color Science, vol 8. Wiley, New York, New York
24. Krizhevsky A, Hinton G et al (2009) Learning multiple layers of features from tiny images
25. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. International journal of computer vision. 115(3):211–252
26. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4700–4708
27. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
28. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings
29. Howard A, Sandler M, Chu G, Chen L–C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V et al (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1314–1324
30. Flynn JR, Ward S, Abich J, Poole D (2013) Image quality assessment using the ssim and the just noticeable difference paradigm. In: International Conference on Engineering Psychology and Cognitive Ergonomics, pp 23–30 . Springer
31. Liu Y, Zhang D, Zhang Q, Han J (2022) Part-object relational visual saliency. IEEE Trans Pattern Anal Mach Intell 44(7):3688–3704
32. Liu Y, Zhang D, Zhang Q, Han J (2021) Integrating part-object relationship and contrast for camouflaged object detection. IEEE Trans Inf Forensics Secur 16:5154–5166
33. Xu S, Gu J, Hua Y, Liu Y (2023) Dktnet: Dual-key transformer network for small object detection. Neurocomputing 525:29–41

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

Authors and Affiliations

Chen-Kuo Chiang[1] · Ying-Dar Lin[2] · Ren-Hung Hwang[3] · Po-Ching Lin[4] · Shih-Ya Chang[4] · Hao-Ting Li[4]

Ying-Dar Lin
ydlin@cs.nctu.edu.tw

Ren-Hung Hwang
rhhwang@nycu.edu.tw

Po-Ching Lin
pclin@cs.ccu.edu.tw

Shih-Ya Chang
otischang.true@gmail.com

Hao-Ting Li
haotingli@csie.io

[1] Computer Science and Information Engineering, Advanced Institute of Manufacturing with High-tech Innovations and Center for Innovative Research on Aging Society (CIRAS), National Chung Cheng University, 62130 Chiayi, Taiwan

[2] Computer Science, National Yang Ming Chiao Tung University, 300093 Hsinchu, Taiwan

[3] College of Artificial Intelligence, National Yang Ming Chiao Tung University, 71150 Tainan, Taiwan

[4] Computer Science and Information Engineering, National Chung Cheng University, 62130 Chiayi, Taiwan