information

MDPI

Article

DIPA: Adversarial Attack on DNNs by Dropping Information and Pixel-Level Attack on Attention

Jing Liu [1], Huailin Liu [1], Pengju Wang [1], Yang Wu [1] and Keqin Li [2,*]

[1] College of Computer Science, Inner Mongolia University, Hohhot 010021, China; liujing@imu.edu.cn (J.L.); huailin.liuzzz@gmail.com (H.L.); 22209004@mail.imu.edu.cn (P.W.); 22109007@mail.imu.edu.cn (Y.W.)
[2] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA
* Correspondence: lik@newpaltz.edu

Abstract: Deep neural networks (DNNs) have shown remarkable performance across a wide range of fields, including image recognition, natural language processing, and speech processing. However, recent studies indicate that DNNs are highly vulnerable to well-crafted adversarial samples, which can cause incorrect classifications and predictions. These samples are so similar to the original ones that they are nearly undetectable by human vision, posing a significant security risk to DNNs in the real world due to the impact of adversarial attacks. Currently, the most common adversarial attack methods explicitly add adversarial perturbations to image samples, often resulting in adversarial samples that are easier to distinguish by humans. To address this issue, we are motivated to develop more effective methods for generating adversarial samples that remain undetectable to human vision. This paper proposes a pixel-level adversarial attack method based on attention mechanism and high-frequency information separation, named DIPA. Specifically, our approach involves constructing an attention suppression loss function and utilizing gradient information to identify and perturb sensitive pixels. By suppressing the model's attention to the correct classes, the neural network is misled to focus on irrelevant classes, leading to incorrect judgments. Unlike previous studies, DIPA enhances the attack of adversarial samples by separating the imperceptible details in image samples to more effectively hide the adversarial perturbation while ensuring a higher attack success rate. Our experimental results demonstrate that under the extreme single-pixel attack scenario, DIPA achieves higher attack success rates for neural network models with various architectures. Furthermore, the visualization results and quantitative metrics illustrate that the DIPA can generate more imperceptible adversarial perturbation.

Keywords: adversarial attack; attention mechanism; high-frequency information; pixel-level attack

Citation: Liu, J.; Liu, H.; Wang, P.; Wu, Y.; Li, K. DIPA: Adversarial Attack on DNNs by Dropping Information and Pixel-Level Attack on Attention. *Information* **2024**, *15*, 391. https://doi.org/10.3390/info15070391

Academic Editors: Eftim Zdravevski, Petre Lameski and Ivan Miguel Pires

Received: 4 June 2024
Revised: 1 July 2024
Accepted: 2 July 2024
Published: 3 July 2024

1. Introduction

Deep learning was proposed by Geoffrey Hinton in 2006 [1], a neural network expert at the University of Toronto. Currently, deep learning is employed significantly in various fields of vital application. These fields include the recognition of images, natural language, speech, and autonomous driving. In the field of image recognition, methods based on deep neural networks have surpassed traditional image processing techniques and even surpassed human processing efficiency [2]. Deep neural networks possess efficient high-level processing and strong learning abilities. However, there are pressing issues that require prompt attention. Recent research reveals that deep neural networks are highly vulnerable to adversarial samples, which result in incorrect decisions. If incorrect decisions occur in application scenarios that heavily depend in deep neural networks, the consequences could be catastrophic. Szegedy et al. first introduced the concept of adversarial samples for image recognition in computer vision [3]. Adversarial samples refer to image samples that closely resemble the original benign samples to the point where they are indistinguishable from the human visual system. However, these samples lead to wrong classification decisions with

a high degree of confidence in the deep neural network model. Attacking a neural network model by creating an adversarial sample technique is called an adversarial attack [4]. Due to the vulnerability of the deep neural network model to adversarial attacks, some of its essential applications are held back by potential security risks. Malicious adversarial attacks deceive automatic driving systems, leading to faulty recognition of traffic signs and, ultimately, causing road accidents. Adversarial samples can also be used to attack face recognition systems, leading to property damage and security challenges. Consequently, developing efficient adversarial attack algorithms will help test the security of deep neural network models and guide the development of adversarial defense mechanisms.

Adversarial attacks are classified into two types based on the access of the attacker to the parameters and structure of the deep neural network model. These types are known as white-box adversarial attacks and black-box adversarial attacks. In the white-box attack, the attacker has complete access to the specific parameters and structure of the neural network model. On the other hand, in the black-box attack, the attacker only has access to the input and output data of the neural network model and cannot access the internal information of the neural network model. Currently, white-box adversarial attacks are the predominant method used to create adversarial samples. Improving white-box attack techniques can also enable them to be executed in a black-box environment. Therefore, this research concentrates on the impact of adversarial attacks in white-box environments. However, traditional white-box adversarial attack methods utilize the L_2 or L_∞ norm distance as a constraint for perturbing the original benign samples. The goal is to introduce minimal adversarial perturbation to generate adversarial samples, thereby causing incorrect classification decisions by the deep neural network model. The existing work has focused on adding adversarial perturbation to the whole image sample, which leads to the unsatisfactory concealing of adversarial perturbation so that human vision can easily distinguish adversarial samples.

This study proposes a pixel-level adversarial attack method for deep neural network models that utilizes an attention mechanism combined with high-frequency information separation. To increase the effectiveness of concealing adversarial perturbation, our method implements the L_0 norm as a constraint on the number of perturbed pixels, which limits the degree of change between the adversarial and original benign samples. Additionally, we discovered that deep neural network models base their classification decisions on image regions of special attention. These attention regions generally contain the primary object information and semantic features of the image sample. As such, we utilize the attention mechanism to locate relevant image regions in the deep neural network model. We survey these areas to find sensitive pixels and apply adversarial perturbation to these pixels to generate adversarial samples. In addition, based on the particularity of the human visual system (HVS), the human eyes can easily recognize the semantic information of the image samples that have lost some information, but recognizing such image samples is still challenging for deep neural networks. So, we also explore the robustness of the neural network model from another perspective. Human vision is not sensitive to high-frequency components containing object edge texture and complex detail information [5]. Therefore, separating the high-frequency components of the image samples and dropping the imperceptible details in the samples enhances the attack of adversarial samples. In this way, it becomes possible to effectively conceal the adversarial perturbation while simultaneously ensuring a higher success rate for the attack. Our method has significance and research value for further revealing the vulnerability of the deep neural network models and guiding to improve robustness. The main contributions can be summarized as follows.

(1) We propose an adversarial attack named DIPA, which is based on an attention mechanism and high-frequency information separation. The adversarial perturbation is generated by attacking the attention of neural network models combined with the separation of high-frequency information from image samples. Simultaneously,

we employ the L_0 norm to limit adversarial perturbation, thereby reinforcing a high attack success rate and effectively concealing it.

(2) We conduct extensive experiments on the ImageNet dataset, setting up two different scenarios. The experiment results show that, compared with the existing AoA method and one-pixel attack method, our method achieves better results on several evaluating metrics.

(3) Finally, we use visualization and quantification to analyze the concealing effect of adversarial disturbance and compare it with many traditional adversarial attack methods, which verify that our method can generate imperceptible adversarial perturbation.

2. Related Work

Currently, the systems deployed based on deep neural network models are more vulnerable to adversarial samples, so the robustness of deep neural networks has become a hot topic [6]. The concept of adversarial samples was initially proposed by Szegedy et al. [3] in their work. Simultaneously, the pioneering white-box adversarial attack method was proposed, using the L-BFGS method with box constraints to calculate the minimal adversarial perturbation, which was subsequently added to the original benign samples, resulting in the generation of adversarial samples. In 2014, Goodfellow et al. [7] proposed an adversarial sample generation method based on the principle of gradient descent, named FGSM. FGSM induces the network to misclassify the generated images by adding increments in the gradient direction, and the gradient can be calculated through the backpropagation algorithm. Kurakin et al. [8] proposed the basic iterative method, named BIM, which perturbs images to obtain better adversarial examples through multiple small steps to increase the loss function of the classifier by optimizing a large step of operation. C&W attacks were proposed by Carlini et al. [9] based on the summary of L-BFGS, FGSM, and JSMA. The algorithm improves on all three methods in the L_0, L_2, and L_∞ norms. Therefore, the success of this type of attack requires that the difference between the adversarial example and the original image be as small as possible. The adversarial example causes the model to misclassify, and the confidence of the misclassified class should be as high as possible. Cheng et al. [10] proposed a prior-guided Bayesian optimization (P-BO) algorithm, which utilizes a surrogate model as a global function prior to black-box adversarial attacks. P-BO models the attack objective using Gaussian processes, and experiments demonstrate its superiority in reducing query numbers and improving attack success rates. Duan et al. [11] proposed a novel adversarial attack method, which uses JPEG compression technology to drop part of the image information in a quantized manner. Moreover, the gradient information of backpropagation is used to optimize the quantization table to reduce the distortion of image samples. Liu et al. [12] proposed an adversarial semantic mask attack framework (ASMA), which can constrain the generated perturbations within local semantic regions, producing adversarial examples with good transferability and stealthiness.

Due to the unknowable constraints of the internal information of deep neural networks, black-box adversarial attacks face greater challenges, which can usually be divided into two categories. Currently, black-box attack methods are mainly divided into query-based methods [13] and transferability-based methods [14]. Chen et al. [15] proposed the attention-based universal adversarial attack (AoA), which is a black-box attack method based on transferability. AoA changes the attention heat map of the original samples to generate adversarial samples. Huang et al. [16] proposed a black-box adversarial attack algorithm based on an evolutionary strategy and attention mechanism. This method fully considers the distribution of gradient update direction in the process of attack, adaptively learns a better search path, improves the efficiency of attack, and combines attention mechanism to eliminate redundant adversarial perturbation. Finally, the imperceptibility of the adversarial sample is improved. M Duan et al. [17] proposed an adversarial samples generation method based on a dual attention mechanism named DAAN. The researchers used spatial and channel attention mechanisms to exploit key regions of the image feature map,

which provide accurate directions for generating effective perturbations. Finally, smooth and tiny perturbations are added to key regions of image samples to generate high-quality adversarial samples. C Lin et al. [18] proposed a black-box adversarial attack targeting sensitive regions by perturbing key pixel locations in an image to generate adversarial examples. To effectively locate key pixels, the literature employs surrogate models and attention heatmap techniques to create highly transferable adversarial examples. H Liu et al. [19] used attention heatmaps as masks combined with low-frequency noise information obtained through random sampling to find adversarial points via binary search. Finally, it conducts local geometric probing near the decision boundary to reduce the sample distance.

In order to ensure the imperceptibility of adversarial perturbation, existing studies have proposed generating adversarial samples from perturbed pixels. Su et al. [20] proposed a one-pixel attack based on a differential evolution algorithm, exploring the attack mode under extreme conditions; that is, perturbing only one pixel in the image to deceive the classifier. The one-pixel attack achieves good attack performance on the adversarial samples generated by perturbing only one pixel on the Kaggle CIFAR-10 dataset. Papernot et al. [21,22] proposed an adversarial attack method called JSMA, which perturbs image samples with the L_0 norm distance as a constraint. This method builds a saliency map based on the forward derivative to reflect which pixels have a greater impact on the image. Selecting the pixel with the largest pixel value in the saliency map as the disturbance point adds a fixed amplitude disturbance, and the process is repeated continuously until the attack is successful. JSMA perturbs numerous pixels and has a high time complexity when generating large-size adversarial samples. So, the attack effect on large-scale datasets such as ImageNet is not ideal. Liu et al. [23] proposed an attack method for generating adversarial samples by perturbing partial pixels, namely PlAA. This method generates adversarial samples by iteratively perturbing sensitive pixels in the attention region of deep neural networks. PlAA is a result of our research in 2022. We combine PlAA with high-frequency information separation technology, namely DIPA, while limiting the number of pixels in the semantic region of the perturbed image samples; then, we drop the imperceptible details in the samples to enhance the attack performance of adversarial samples. Finally, the adversarial samples produced by DIPA exhibit improved effectiveness in concealing the adversarial perturbation while maintaining a higher rate of success in adversarial attacks. P N Williams et al. [24] proposed a bi-objective optimization adversarial attack algorithm based on L_0 constraint. By introducing the L_0 constraint to limit adversarial perturbations to a few pixels, the algorithm enhances the imperceptibility of adversarial perturbations.

3. Methodology

The primary objective of this research is to address the inefficiency of traditional adversarial attack methods in concealing adversarial perturbation. We introduce a pixel-level adversarial attack algorithm based on the attention mechanism and high-frequency information separation to counteract this issue. Our proposed method utilizes the attention heat map generated by the deep neural network's attention mechanism. By designing the loss function that diminishes the attention of the neural network, we iteratively perturb single pixels to suppress the network's attention towards the correct class. The outcome is an erroneous classification decision, as the model redirects its attention towards the irrelevant class. The human visual system has the ability to quickly locate and identify the primary object in an image while also interpreting its semantic information. However, the details and complexities of the foreground and background are often overlooked by human visual perception. Within the frequency domain, the low-frequency component of an image typically comprises the main semantic information and the basic structure of the object in question, whereas the high-frequency component generally encompasses the target object's edge details and complex background information. To separate high-frequency information from the image, we employ the discrete wavelet transform frequency domain processing technology. The justification for using the discrete wavelet transform (DWT) for

high-frequency information separation and its suitability for the study can be detailed as follows. DWT provides a multi-resolution analysis by decomposing a signal into different frequency components while maintaining temporal localization. DWT is versatile and can be adapted to different types of signals by choosing appropriate wavelet functions (such as Haar, Daubechies, etc.). DWT has been widely used and validated in various domains, such as signal processing, image processing, and data compression, demonstrating its robustness and effectiveness in high-frequency information separation. The study can rely on this proven track record to ensure that the methodology is sound and has a high likelihood of success in achieving the desired outcomes. Additionally, the number of perturbed pixels is limited to ensure high-quality adversarial samples are generated. This paper presents a pixel-level adversarial sample generation method based on attention mechanism and high-frequency information separation, which includes four key stages: discrete wavelet transform for high-frequency information separation, attention heat map generation and application, attention suppression loss function design, and pixel-level adversarial sample generation. The specific frame design is shown in Figure 1.

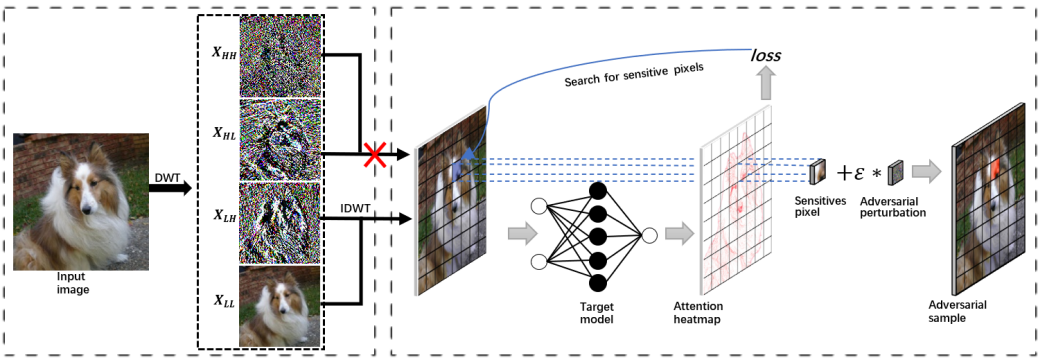

Figure 1. Workflow of our DIPA method, which illustrates the process of high-frequency information separation (**left**) and pixel-level attack on attention (**right**).

Firstly, the image samples are decomposed into the frequency domain using the discrete wavelet transform. High-frequency information on vertical and diagonal edge features is dropped during image reconstruction, perturbing the resulting image samples at the pixel level. Secondly, an attention heat map is generated using a deep neural network's attention mechanisms to highlight specific regions of the image. The attention suppression loss function is then designed based on this map, which allows the neural network to potentially make incorrect classification decisions by suppressing its attention toward the correct image class. The DIPA method generates adversarial samples by searching for and perturbing sensitive pixels based on the gradient information of the loss function to the original benign samples. The method uses the L_0 norm as a constraint to limit the number of perturbed pixels. This approach is effective at increasing attack success rates and concealing adversarial perturbation.

3.1. High-Frequency Information Separation Based on Discrete Wavelet Transform

Based on current research on the human visual system, it is evident that humans are more sensitive to object structures and smooth areas of an image. Adversarial perturbations are, therefore, generally observable, but intricate texture details, including object edges and complex background information, are less noticeable. As shown in Figure 2, x represents the input image, and $\bar{x}$ is the image sample reconstructed using low-frequency components (LL) and horizontal edge features (LH), which have the same basic shape and resolution as the original image x. Image distortion in the dense fishnet background in part A is less noticeable to the human eye compared to distortion on smooth areas and subjects in part B. This prompts us to separate the high-frequency components representing complex background information and to attack the neural network model by dropping irrelevant information and perturbing sensitive pixels. As a time-frequency analysis tool, discrete

wavelet transform can convert image samples from the spatial domain to the frequency domain and decompose samples into a low-frequency component and three high-frequency components, namely X_{LL}, X_{LH}, X_{HL}, and X_{HH}. The calculation formulas are shown in (1) and (2).

$$X_{LL} = LXL^T, X_{LH} = HXL^T, \tag{1}$$

$$X_{HL} = LXH^T, X_{HH} = HXH^T, \tag{2}$$

where L and H represent the low-pass filter and high-pass filter of the orthogonal wavelet, respectively. This paper uses the Haar wavelet basis function, and the coefficients of the low-pass filter are shown in Formula (3).

$$\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, \tag{3}$$

where L contains the low-pass decomposition filter coefficient and low-pass reconstruction filter coefficient. High-pass filter coefficients are shown in Formula (4).

$$\begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}. \tag{4}$$

The high-pass filter also includes high-pass decomposition filter coefficients and high-pass reconstruction filter coefficients. The two-dimensional Haar discrete wavelet transform first uses a low-pass filter and a high-pass filter to perform column filtering on the image. In order to ensure the computational simplicity of the Haar discrete wavelet transform, the frequency component will be downsampled with a frequency of 2 after each filtering process. Then, the low-pass filter and high-pass filter are used to process each frequency component again, and the image sample is decomposed into four frequency sub-bands. We use X_{LL} and X_{LH} to reconstruct the image and change downsampling to upsampling with a frequency of 2. At the same time, considering that only two high-frequency components, X_{HL} and X_{HH}, are dropped, the image will not be seriously distorted due to the loss of too much detailed information. The image reconstruction formula is shown in (5).

$$\begin{aligned} \bar{X} &= L^T X_{LL} L + H^T X_{LH} L \\ &= L^T (LXL^T) L + H^T (HXL^T) L. \end{aligned} \tag{5}$$

Facilitating image decomposition and reconstruction, discrete wavelet transform can preserve critical image information. A high-quality adversarial sample can then be produced by combining the perturbation of sensitive pixels with the separation of high-frequency information.

3.2. Designing Attention Suppression Loss Function

The suppression the loss function is designed with the intention of reducing the attention of deep neural networks on the correct class of an image sample. By doing so, the model's attention towards other irrelevant classes gradually surpasses the attention towards the correct class. Consequently, this leads to erroneous classification decisions made by the deep neural network models. Hence, our approach leverages the softmax gradient LRP (SGLRP) method to generate an attention heat map that effectively differentiates the attention towards the target class from other irrelevant classes [25]. Furthermore, we compute the gradient of the attention suppression loss function with respect to the input image sample. The input image, which has been separated from some high-frequency information by discrete wavelet transform, is utilized to calculate the attention heat map. The attention heat map generated by the deep neural network models of different architectures is shown in Figure 3.

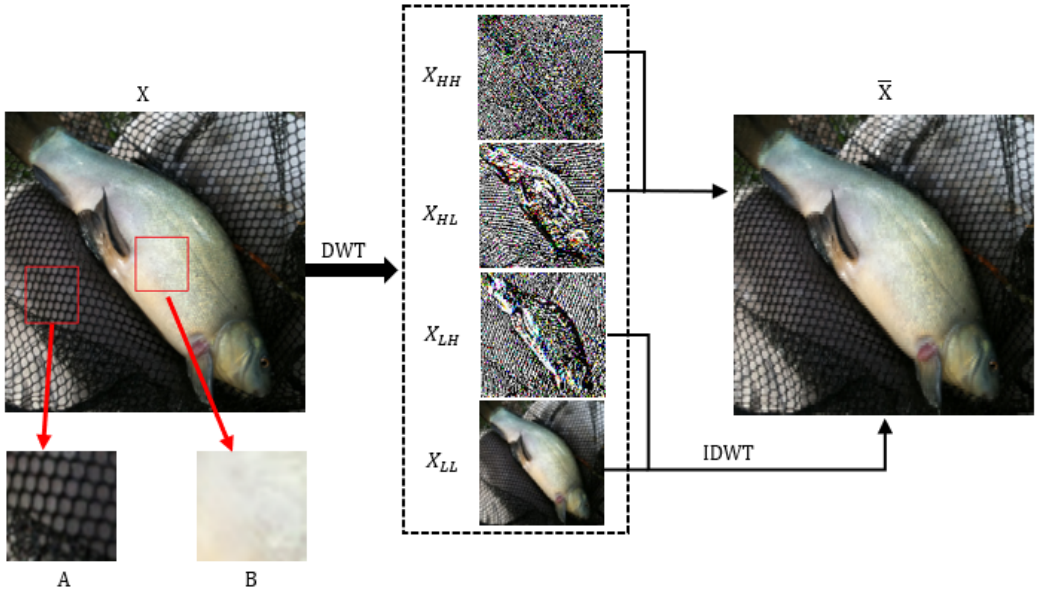

Figure 2. Illustration of high-frequency information separation. The complex background information (fishnet) dropped by DIPA is difficult for human vision to perceive. Furthermore, restricting adversarial perturbation on object structure and smooth region can improve the perceptual quality of adversarial samples.

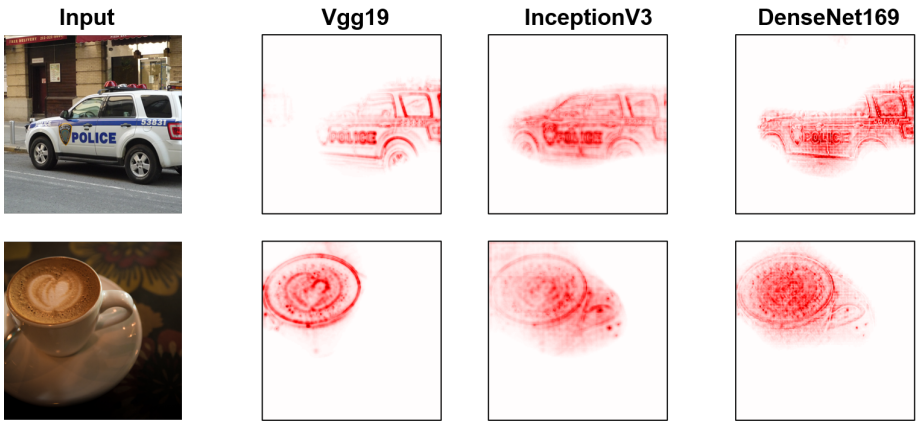

Figure 3. Attention heat maps for DNNs. The pixel-wise heat maps show how the input contributes to the prediction. Deep neural network models with different architectures have similar attention heat maps.

We set x_{ori} as a benign input sample and y_{ori} as its corresponding class label. The attention heat map of the image sample is represented by $h(x_{ori}, y_{ori})$, which has the same matrix dimension as x_{ori}. The attention suppression loss function *Loss* comprises two loss functions: the logarithmic suppression loss function L_{log} and the cross-entropy loss function L_c. The L_{log} is shown in Equation (6).

$$L_{\log}(x) = \log(\|h(x, y_{\text{ori}})\|_1). \tag{6}$$

The logarithmic suppression loss function serves to diminish the deep neural network model's attention towards the correct class of image sample, leading to erroneous classification decisions. The L_c is shown in Equation (7).

$$L_c(x) = -\sum_{i=1}^{n} p(x_i) \log(q(x_i)). \tag{7}$$

The cross-entropy loss function is used to increase the confidence of adversarial samples, which ultimately leads to deep neural networks classifying these samples into incorrect classes with high confidence. The $Loss(x)$ is shown in Equation (8).

$$Loss(x) = L_{\log}(x) + \alpha L_c(p, q). \tag{8}$$

The attention suppression loss functions suppress DNNs' attention to the correct classes, as shown in Figure 4. The parameter α is the weight parameter between the logarithmic suppression loss function L_{log} and the cross-entropy loss function L_c. We set the initial value of α to $\alpha = 1000$ because the ImageNet dataset used in the experiments in this paper has a total of 1000 classes, which makes the two components of the suppression loss function have similar variances for different inputs.

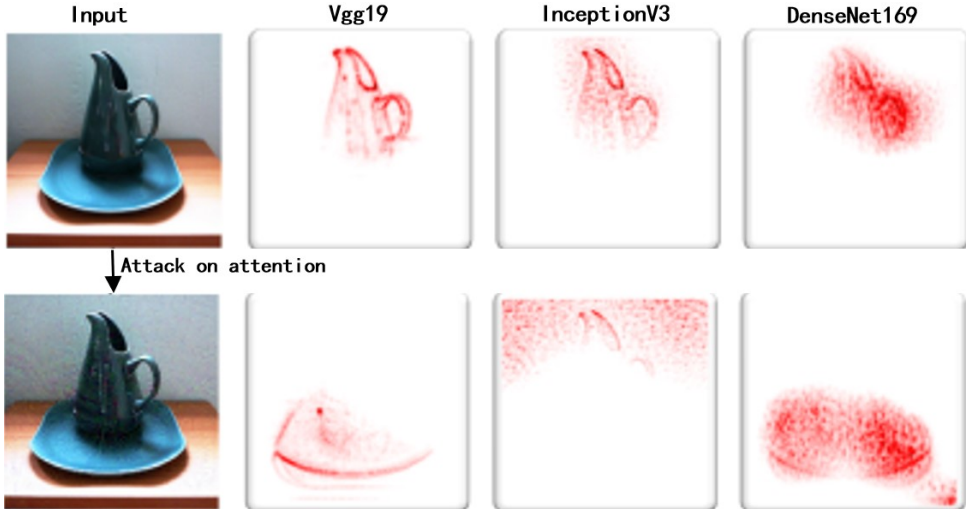

Figure 4. Attack on attention for DNNS. The attention suppression loss function distracts the attention from the correct to irrelevant regions and similar distraction could be observed for different networks.

3.3. Pixel-Level Attack Algorithm Based on Attention Mechanism

Generating adversarial samples is the core step of the adversarial attack algorithm. The main idea of our method is to minimize the attention suppression loss function $Loss(x)$, guided by the gradient information of the loss function to the original benign samples, and iteratively search and perturb sensitive pixels to generate adversarial samples. First, the SGLRP method is used to calculate the attention heat map of the deep neural network models for the input image sample, construct the attention suppression loss function, and calculate the gradient information of the suppression loss function $Loss(x)$ to the input sample x. By obtaining a gradient matrix M that matches the dimensions of the input sample x, we can assess the extent to which each pixel can suppress the attention of the deep neural networks. Each element in the gradient matrix M quantifies the degree of suppression for the corresponding pixel. Hence, based on the gradient matrix M, we identify the pixel with the highest gradient value, as it is deemed to have a substantial influence in suppressing the deep neural network model's attention towards the correct class. Formally, let $x^0_{adv} = x_{ori}$, and the gradient matrix M is calculated as shown in Formula (9).

$$M = \frac{\partial Loss(x^k_{adv})}{\partial x^k_{adv}}. \tag{9}$$

Specifically, the greater the element value in the gradient matrix M, the stronger the effect on suppressing attention. To account for the fact that each pixel value in an image sample is a composition of the brightness values from the three RGB channels, it becomes essential to compute the sum of gradient values across these channels. This aggregation results in the formation of the gradient sum matrix M_{sum}. By evaluating this matrix, we can

identify the pixel with the highest gradient value and its corresponding impact on attention suppression within the neural network model. Subsequently, utilizing the information from M_{sum}, we identify the pixel with the highest gradient value. This particular pixel is then targeted for perturbation, which enables us to generate the adversarial samples. The gradient sum matrix M_{sum} and the disturbance process are shown in Formulas (10)–(12).

$$M_{sum} = sum(M, reduction\_indices = 2), \tag{10}$$

$$index = argmax(M_{sum}), \tag{11}$$

$$x_{adv}^{(k+1)} = x_{adv}^{k}[index] - \varepsilon * M[index]. \tag{12}$$

To indicate the summation of the gradient matrix M along the rows, we have set the parameter $reduction\_indices = 2$ in our specification. In order to perturb the sensitive pixels, we assign the parameter ϵ a specific value, which represents the intensity of the perturbation. Given that we perturb only one pixel in each iteration, the resulting distortion to the image sample is minimal. Consequently, there is no need to enforce any limitations on the strength of the perturbation. The procedure of attack is summarized in Algorithm 1.

Algorithm 1 Adversarial Attack.

Require:
Origin sample (x_{ori}, y_{ori});
Discrete wavelet transform (DWT);
Inverse discrete wavelet transform $(IDWT)$;
Loss $Loss(x)$;
Iterator threshold θ;
Attack step length ϵ;
Ensure:
Adversarial sample x_{adv};

$x_{LL}, x_{LH}, x_{HL}, x_{HH} = DWT(x_{ori})$;
$\bar{x} = IDWT(x_{LL}, x_{LH})$;
$x_{adv}^{0} = \bar{x}$;
set $k = 0$;
while $k < \theta$ **do**
  $M = \frac{\partial Loss(x_{adv}^{k})}{\partial x_{adv}^{k}}$;
  $M_{sum} = sum(M, reduction\_indices = 2)$;
  $index = argmax(M_{sum})$;
  $x_{adv}^{k+1} = x_{adv}^{k}[index[0:2]] - \epsilon * M[index[0:2]]$;
  $k = k + 1$;
  if $F(x_{adv}^{k}) \neq y_{ori}$ **then**
    break;
  end if
end while
return x_{adv}^{k};

Our method mainly adds pixel-level adversarial perturbation to image samples which drop some high-frequency components, and emphasizing the pixel with the largest gradient value has a significant impact on suppressing the attention of the neural network model. Therefore, adversarial samples can be generated by iteratively perturbing only a few sensitive pixels with the maximum gradient value. Additionally, since the adversarial samples have only a few pixel changes in the main part of the image compared with the original samples, moreover the dropped part of the information is mainly related to the edge details and complex textures that are not easy to perceive. This is why the overall distortion of the image sample is tiny, and human vision can hardly observe the difference between the adversarial sample and the original sample. Our adversarial attack method

improves the search efficiency of sensitive pixels, further confirming the vulnerability of deep neural networks.

4. Experiments

In this section, we present the experimental process, results, and analysis of our attack method. We conducted experiments in two different scenarios: the extreme single-pixel attack scenario and the white-box attack scenario. We validated the effectiveness and advantages of the DIPA method through two experimental scenarios. We also utilized cutting-edge perceptual similarity measurement technology to assess the perceptual quality of adversarial samples.

4.1. Experiment Setup

The experimental dataset is the validation set of the ImageNet dataset, which has a total of 1000 classes and 50,000 image samples. We chose the ImageNet dataset because it is widely used in current research on image classification. Compared to other datasets like CIFAR-10 or CIFAR-100, ImageNet offers samples with higher resolution and a larger number of categories. This allows for effective validation of our method's efficacy. Although transformer-based models have good performance in computer vision, the actual application is still dominated by CNN-based models [26]. In this experiment, eight well-trained deep neural network models in Keras applications [27] were selected to attack and verify the effectiveness of the method, namely VGG19 and VGG16 [28], InceptionV3 [29], ResNet50 and ResNet152 [30], and DenseNet169, DenseNet121, and DenseNet201 [31]. We preprocessed image samples using Keras preprocessing, which entailed central cropping and resizing to a resolution of 224x224. Initially, the dataset was filtered for each model by purging image samples that have been incorrectly classified. That is, subsequent attacked image samples can be correctly classified by the depth neural network model. We filtered 1000 image samples for each neural network model for attack testing. In addition, the number of iterations was $\theta = 20$, and the disturbance intensity was set to $\epsilon = 155$. The number of iterations θ was used to limit the number of perturbed pixels, and the intensity of disturbance ϵ was used to limit the magnitude of the adversarial perturbations applied to the pixels, preventing noticeable disturbance in the image samples. Through experimental validation and by referring to related literature, setting the number of iterations to $\theta = 20$ and the perturbation strength to $\epsilon = 155$ can achieve an optimal balance between attack success rate and perceptual quality.

4.2. Evaluation Metrics

In terms of attack performance, our focus lies mainly on the attack success rate and the perceptual quality of adversarial samples. Accordingly, we propose six evaluation metrics to assess the effectiveness of the DIPA method.

1. Average root mean square error ($AvgRMSE$):

$$AvgRMSE = \frac{\sum_{i=1}^{N_{Att\_suc}} \sqrt{\left\| x_{adv}^{i} - x_{ori}^{i} \right\|_2^2 / N_p}}{N_{Att\_suc}}. \tag{13}$$

   In the experiment, the deep neural network models generate multiple adversarial samples, which we evaluate using the average root mean square error ($AvgRMSE$) to determine the degree of change. Where N_{Att_suc} represents the number of adversarial samples generated by our method, and N_p represents the total number of pixels in the image sample. The $AvgRMSE$ metric is shown in Equation (13).

2. Attack success rate (ASR):

$$ASR = \frac{N_{Att\_suc}}{N}. \tag{14}$$

   Following the generation of adversarial samples, we input them into the model to gauge their capacity to deceive. Where N_{Att_suc} represents the number of adversarial

samples generated by our method that successfully attacks the target model, and variable N denotes the total number of image samples. The ASR metric is shown in Equation (14).

3. Average confidence ($AvgConfidence$):

$$AvgConfidence = \frac{\sum_{i=1}^{N_{Att\_suc}} p_i}{N_{Att\_suc}}, \tag{15}$$

where N_{Att_suc} represents the number of adversarial samples generated by our method, and P_i represents the confidence when adversarial sample i is misclassified by the neural network model. As shown in Equation (15), the $AvgConfidence$ metric refers to the average confidence when the neural network models misclassify all adversarial samples.

4. Time complexity ($AvgTime$):

$$AvgTime = \frac{\sum_{i=1}^{N_{Att\_suc}} T_i}{N_{Att\_suc}}, \tag{16}$$

where N_{Att_suc} represents the number of adversarial samples generated by our method, and T_i is the time spent in generating adversarial sample i. As shown in Equation (16), the $AvgTime$ metric refers to the average time complexity of our method to generate adversarial samples.

5. Number of disturbed pixels ($AvgPixels$):

$$AvgPixels = \frac{\sum_{j=1}^{N_{Att\_suc}} Pix_j}{N_{Att\_suc}}, \tag{17}$$

where N_{Att_suc} represents the number of adversarial samples generated by our method, and Pix_j represents the number of pixels disturbed by our method when generating adversarial sample j in the white-box attack scenario. $AvgPixels$ is shown in Equation (17).

6. Learning-based perceptual similarity metrics (Lpips): Currently, the most common method to measure the similarity between two image samples is based on distance, such as SSIM [32] and FSIM [33] based on L_2 Euclidean distance, etc. These methods use a simple distance function to calculate the similarity directly. However, humans can easily and quickly assess the perceptual similarity between two images, but the process is highly complex. The method based on distance measurement does not consider the details of human perception and cannot fit well with the human perceptual similarity between two images. Therefore, we choose a learn-based perceptual similarity measurement (Lpips) [34] method that aligns with human perception judgment. The Lpips method visualizes the degree of change between the adversarial sample and the original sample by utilizing a perceptual distance space map. It also quantifies the concealing effect of adversarial perturbation using the Lpips perceptual loss metric. Thus, the advantages and characteristics of the DIPA method can be verified more reasonably.

4.3. Results Analysis in Single-Pixel Attack Scenario

In this section, we evaluate the efficiency and effectiveness of a single-pixel attack scenario, which generates adversarial samples by perturbing only one pixel. The aim of this experiment is to investigate whether attention mechanisms can accurately search for sensitive pixels and to determine whether disturbing sensitive pixels can attack the deep neural network models. We conducted a comparison between our experimental results and the existing one-pixel attack methods [9]. The evaluation metrics for the experimental results of the one-pixel attack method are presented in Table 1. To validate the effectiveness

of the DIPA method, we performed experiments on neural network models with diverse architectures. The experimental results of the single-pixel attack are provided in Table 2.

Table 1. Experimental results of the one-pixel attack.

Metrics Victim	AlexNet_BVLC
ASR	16.04%
AvgConfidence	22.91
AvgRMSE	14.32
AvgTime (s)	-

Table 2. Experimental results of DIPA on different network architectures.

Metrics Victim	AlexNet_BVLC	VGG19	VGG16	IncV3	RN50	RN152	DN121	DN169	DN201
ASR	43.2%	50.2%	43.5%	56.8%	16.7%	23.3%	13.1%	12.4%	15.2%
AvgConfidence	69.552	68.987	69.650	20.833	45.710	55.465	45.461	40.979	49.898
AvgRMSE	16.437	16.534	16.310	12.433	18.434	18.839	18.116	18.629	18.977
AvgTime (s)	1.4	2.4	1.5	6.0	6.4	14.8	35.7	44.6	48.1

It can be seen from Table 1 that the existing one-pixel attack methods mainly use differential evolution algorithms to search for sensitive pixels. They have achieved an attack success rate of only 16.04% on large-size ImageNet datasets. The reason may be that large-size image samples increase the search space for sensitive pixels, making one-pixel attack methods inaccurate in searching for sensitive pixels, resulting in an unsatisfactory attack success rate on large-size image samples.

By comparing the results presented in Table 2 with those in Table 1, it can be seen that the DIPA demonstrates significant enhancements in both the ASR and $AvgConfidence$ metric when applied to large image samples on the same dataset and neural network models. Specifically, in the AlexNet_BVLC model, the DIPA method improves ASR and $AvgConfidence$ metrics by 27.16% and 46.642, respectively. Additionally, Table 2 provides insights into the impact of DNN depth on the ASR and $AvgTime$ of single-pixel attacks. As the depth of the network model increases, the ASR gradually decreases, but the $AvgTime$ gradually increases. This indicates that deep neural networks incur a higher time cost, and the attack success rate of the single-pixel attack is relatively poor. However, shallow neural network models can achieve an attack success rate of more than 40% in just a few seconds. On average, all deep neural network models misclassify adversarial examples generated by single-pixel attacks with high confidence. It is worth mentioning that the experimental results verify the effectiveness and efficiency of the DIPA method in a single pixel attack scenario.

4.4. Results Analysis in the White-Box Attack Scenario

This section assesses the efficacy and efficiency of our method in the context of white-box attacks. Firstly, the high-frequency information of the input image is separated, and then sensitive pixel points are iteratively selected and perturbed using the gradient information of the target neural network model. Our method compares the existing adversarial attack on attention methods (AoA) [10]. AoA attacks the attention of neural network models and achieves a stronger attack effect. The experimental results are shown in Table 3. As shown in Table 3, from the overall perspective, the DIPA method achieved a comparable attack success rate to that of the AoA method. Most networks exhibit an attack success rate exceeding 90%.

Table 3. Attack success rate of white-box attacks on different network architectures.

Metrics Victim	VGG19	VGG16	IncV3	RN50	RN152	DN121	DN169	DN201
AoA	99.99%	99.85%	89.84%	93.94%	86.78%	96.14%	94.09%	93.44%
DIPA	98.8%	98.5%	96.4%	97.6%	95.6%	84.7%	88.4%	93.2%

In addition, except for Inception-V3, all network models can misclassify adversarial samples with high confidence. The $AvgConfidence$ metric for Inception-V3 is only 21.054, and we speculate that it may be related to the network architecture of the Inception-V3 neural network model. It can be seen from Table 3 that the time complexity is affected by network depth. On average, DIPA is capable of attacking shallow neural network models within a few seconds. For instance, the attack duration for VGG19 is approximately 3.3 s, VGG16 takes around 3.5 s, and Inception-V3 requires approximately 8.2 s. Deep neural network models with more network layers require more time to attack. For example, ResNet50 is 18.9 s, ResNet152 is 32.1 s, and DenseNet121 is 38.5 s, and the time complexity of the DenseNet169 and DenseNet201 neural network models is approximately one minute. The average root mean square error serves as an indicator of the degree of modification in the adversarial sample relative to the original sample. The Inception-V3 neural network model has the lowest $AvgRMSE$ of 18.742. A higher $AvgRMSE$ indicates that generating an adversarial sample requires the disturbance of a more significant number of pixels, resulting in more significant changes when compared to the original sample. According to Table 4, the average number of pixels modified to generate adversarial samples for all networks is ten or fewer. This indicates that the DIPA method can still generate tiny adversarial perturbations and more realistic adversarial samples on ImageNet datasets with complex features.

Table 4. Evaluation metrics of DIPA on different network architectures.

Metrics Victim	VGG19	VGG16	IncV3	RN50	RN152	DN121	DN169	DN201
AvgPixels	2.415	2.561	2.318	5.231	4.448	8.372	8.415	6.483
AvgConfidence	69.940	70.260	21.054	43.977	58.185	43.497	47.696	43.910
AvgRMSE	20.143	20.158	18.742	28.970	30.744	29.907	32.703	25.621
AvgTime (s)	3.3	3.5	8.2	18.9	32.1	38.5	54.6	52.5

4.5. Results Analysis in Perceptual Quality

Most methods of adversarial attacks are based on the L_2 or L_∞ norms and are used to limit the intensity of adversarial perturbation, but in order to achieve a higher attack success rate, the concealing effect of adversarial perturbation is not ideal. Existing AoA methods typically impose an overall constraint on the cumulative modification intensity as the primary constraint condition. These methods perturb all pixels of the image to generate adversarial samples. In contrast, our method deviates from traditional attack on attention methods. Our approach focuses on selectively modifying only a small number of pixels within a short time frame without restricting the intensity of the modifications. As a result, our method achieves minimal overall image distortion while generating adversarial samples. Moreover, in contrast to traditional adversarial attack methods that add perturbation to image samples, our method attacks neural network models by dropping some irrelevant image information, which can more effectively hide adversarial perturbation.

4.5.1. Perceptual Quality Visualization Experiment

The adversarial examples generated by our method are shown in Figure 5.

Figure 5. Illustration of adversarial samples. (**a**) shows the original benign sample, (**b**) shows the adversarial sample generated by the AoA method, and (**c**) shows the adversarial sample generated by DIPA. The prediction classes of the ResNet50 and corresponding confidence are provided.

Figure 5a,b shows the benign and the adversarial samples generated by the AoA method, respectively. Figure 5c shows the adversarial samples generated by our method. The prediction class of the neural network model and the corresponding confidence are provided below the image. Upon comparing the adversarial images, it becomes apparent that the ones generated by AoA exhibit prominent wavy patterns that span the entire image, while our method generates adversarial images that are nearly indistinguishable from the original samples. Additionally, relying solely on visual perception is insufficient to demonstrate the effectiveness of our method in concealing adversarial perturbation. Thus, we employ the Lpips perceptual loss as a quantitative perceptual metric to assess the perceptual quality of adversarial samples. The way to measure the similarity between two images is similar to the way humans visual, so our method can more effectively quantify the concealing effect of adversarial perturbation. The lower the Lpips perceived loss value, the higher the perceived quality in adversarial samples.

Figure 6 illustrates the perceptual distance space map generated using the Lpips method. The perceptual loss computed through the Lpips method is presented below the image. Our method outperforms the AoA method, as demonstrated by the lower perceptual loss (0.073 compared to 0.185). This indicates that compared to the AoA method, the adversarial samples generated by our method are more perceptual, consistent with benign samples. The Lpips perception metric provides additional evidence supporting the effectiveness of our method in concealing adversarial perturbations. As shown in Figure 6a, it can be observed that compared to benign samples, the adversarial samples of AoA have significant differences in the main semantic objects in the image. This includes perturbation made to the main object, which is the area of special attention to human vision. However, as shown in Figure 6b, the difference between the adversarial samples generated by our method and the benign samples is mainly distributed on the irrelevant background, the perturbation intensity is low, and only a few pixels in the main semantic object area change. In other words, our method is based on the visual characteristics of human vision and comprehensively considers two aspects. Firstly, because human vision always pays attention to the structural part of the object when judging the semantic information of the image, and the deep neural network models also make classification decisions based on the attention region. Therefore, we limit the adversarial perturbation in the main semantic object region and accurately search and perturb a few of the sensitive pixels based on the attention mechanism. Secondly, human vision is not sensitive to the object edge details and complex background information of the image. We drop high-frequency information instead of adding perturbation when attacking neural network models to achieve a balance between attack success rates and the quality of adversarial samples.

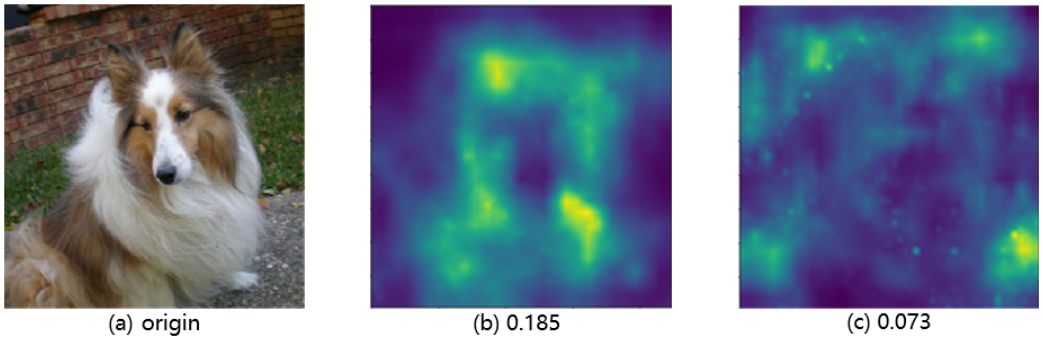

Figure 6. Lpips Perceptual distance space map. The original benign samples are compared with the adversarial samples on pixel-by-pixel basis to generate perceptual distance space map, which shows the distribution of the perturbed pixels, and the pixel brightness shows perturbation intensity.

In addition, the mechanism of separating high-frequency information is to generate adversarial samples by removing imperceptible details from image samples. The experiment also uses K-means clustering to analyze the color depth of image samples. As depicted in Figure 7, adversarial samples generated by AoA require more storage space due to their larger image size and additional information, including more colors. For example, the image sample labels are Guenon and Sheepdog, the image sizes are 101 KB and 94.9 KB, and the color components are 205 colors and 194 colors, respectively. The labels of the adversarial samples generated by the AoA method are Macaque and Hound. The image sizes are increased to 110 KB and 99.8 KB, respectively, and the image sizes are increased by 9% and 5% compared with the original samples. In contrast, the adversarial samples generated by separating high-frequency information contain fewer colors and are labeled Howler and Samoyed, respectively. The color composition is reduced to 97 colors and 82 colors, representing a size reduction of 39% and 64% compared to the original samples. The method of removing detailed information from image samples in the frequency domain appears to be a promising technique for adversarial attacks.

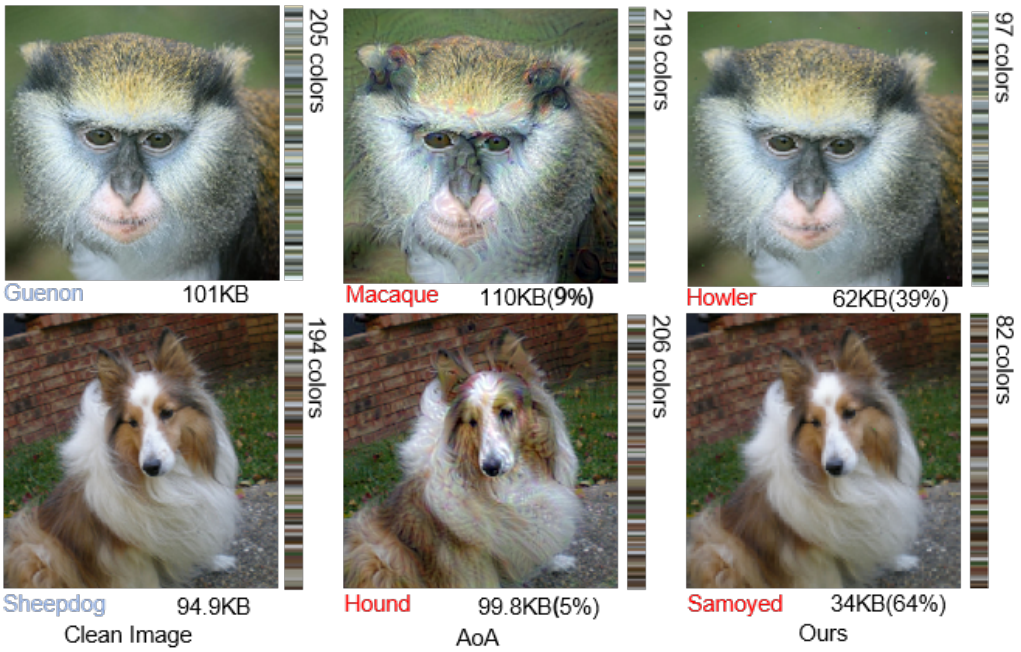

Figure 7. Illustration of separating high-frequency information. Compared to the clean images, the adversarial images generated by DIPA have fewer details composed of fewer colors, decreasing in size (by 39% and 64%).

4.5.2. Comparison with Traditional Adversarial Attack Methods

In addition, we assess and compare the strength and perceived quality of the adversarial samples generated by several traditional adversarial attack methods. The parameters of the traditional adversarial attack method are set as follows. The learning rate of the Adam optimizer for the C&W adversarial attack method is set to 0.01, the maximum disturbance limit ϵ of BIM, PGD [35], and AutoAttack (AA) [36] methods under L_∞ norm constraints is set to $8/255$, the iteration step for each round is set to $\alpha = 1/255$, and the disturbance strength of FGSM method is $eps = 0.1$. Our experiment conducted validation and comparison on seven neural network models, including three neural network models with different architectures and multiple neural network models with the same architecture and different depths. The accuracy and parameter number of image classification tasks on the ImageNet dataset are shown in Table 5.

Table 5. Pretraining model accuracy information.

Model	Acc Top1	Acc Top5	Params
VGG16	73.360%	91.516%	138.4 M
VGG19	74.218%	91.842%	143.7 M
ResNet50	81.198%	95.340%	25.0 M
ResNet152	82.284%	96.002%	60.2 M
DenseNet121	74.434%	91.972%	8.0 M
DenseNet169	75.600%	92.806%	14.1 M
DenseNet201	76.896%	93.37%	20.0 M

Models with more sophisticated architectures, such as deeper networks or those with attention mechanisms, generally performed better. These models are better at capturing intricate patterns and can be more resistant to adversarial attacks. Simpler models, or those with fewer layers, showed decreased performance as they are less capable of handling complex data perturbations introduced by DIPA. Some models exhibited a peculiar increase in $AvgConfidence$, indicating that adversarial examples were not just misclassified but misclassified with high confidence. This suggests that DIPA effectively exploits specific vulnerabilities in these models. While some models showed a high $AvgRMSE$, indicating significant changes from the original samples, others had a lower $AvgRMSE$, suggesting subtler but effective perturbations. This variability points to differences in how models perceive and react to adversarial noise.

This experiment verifies the performance of five attack methods under four metrics, including iteration times, time complexity, attack success rate, and Lpips perceived loss. Each model is attacked by the corresponding adversarial attack method to generate 200 adversarial samples, and the attack success rate and average Lpips loss are calculated. As shown in Table 6. Our method outperforms five traditional adversarial attack methods according to the Lipis metric. These results demonstrate that the adversarial samples generated using our proposed attack method are visually more similar to the original and induce fewer changes in the feature space. DIPA has the smallest Lpips value while achieving a high attack success rate. DIPA achieves a balance between a high attack success rate and a low imperceptibility.

The C&W adversarial attack method is an optimization-based technique that iteratively optimizes the target loss function to search for the smallest perturbation required to generate adversarial samples. We set the number of iterations of the C&W attack method to 1000 in this experiment. Although the C&W method has higher time complexity and better conceals adversarial perturbation under the l_∞ norm limit, perturbation waves in smooth areas of an image can still be easily observed.

Table 6. Comparison of experimental results.

Model	Attack	Iteration	Run Time (s)	ASR	Lpips
VGG16	FGSM	1	22	96.2%	0.306
	BIM	10	243	94.0%	0.089
	PGD	10	56	98.3%	0.129
	C&W	1000	≥10,000	99.2%	0.394
	AA	100	62	99.5%	0.394
	Ours	20	292	98.5%	0.060
VGG19	FGSM	1	24	94.1%	0.301
	BIM	10	255	94.3%	0.091
	PGD	10	66	97.9%	0.128
	C&W	1000	≥10,000	99.3%	0.395
	AA	100	72	99.2%	0.395
	Ours	20	300	98.8%	0.062
ResNet50	FGSM	1	85	94.1%	0.304
	BIM	10	248	92.6%	0.091
	PGD	10	187	99.9%	0.107
	C&W	1000	≥10,000	98.8%	0.390
	AA	100	74	96.8%	0.396
	Ours	20	1405	97.6%	0.063
ResNet152	FGSM	1	67	94.2%	0.308
	BIM	10	320	91.7%	0.096
	PGD	10	257	99.7%	0.126
	C&W	1000	≥10,000	97.3%	0.397
	AA	100	192	97.1%	0.396
	Ours	20	≥1000	95.6%	0.068
DenseNet121	FGSM	1	63	98.5%	0.307
	BIM	10	264	92.4%	0.094
	PGD	10	167	99.4%	0.12
	C&W	1000	≥10,000	98.3%	0.402
	AA	100	126	97.3%	0.387
	Ours	20	≥5000	90.7%	0.070
DenseNet169	FGSM	1	75	96.7%	0.308
	BIM	10	229	90.8%	0.092
	PGD	10	231	99.7%	0.126
	C&W	1000	≥10,000	96.2%	0.407
	AA	100	182	97.1%	0.384
	Ours	20	≥5000	89.4%	0.068
DenseNet201	FGSM	1	131	94.0%	0.304
	BIM	10	330	93.8%	0.092
	PGD	10	273	99.4	0.126
	C&W	1000	≥10,000	95.4	0.408
	AA	100	246	98.5	0.389
	Ours	20	≥5000	93.2	0.069

As shown in Figure 8, compared with the perceptual distance space map, the C&W method adds perturbation to the entire image, resulting in slight distortion to the image sample. According to the perceptual distance space map of our method, there may be slight distortion in the edge details and complex textures of image objects due to the separation of high-frequency information. However, the human visual system is not sensitive to edge distortion and does not pay attention to irrelevant details. Preserving object structure and low-frequency components that are sensitive to human visual perception makes the DIPA method ideal for concealing adversarial perturbation. In addition, DIPA adds fewer adversarial perturbations to generate adversarial samples, and the adversarial perturbations are mainly distributed in the object edge and the background.

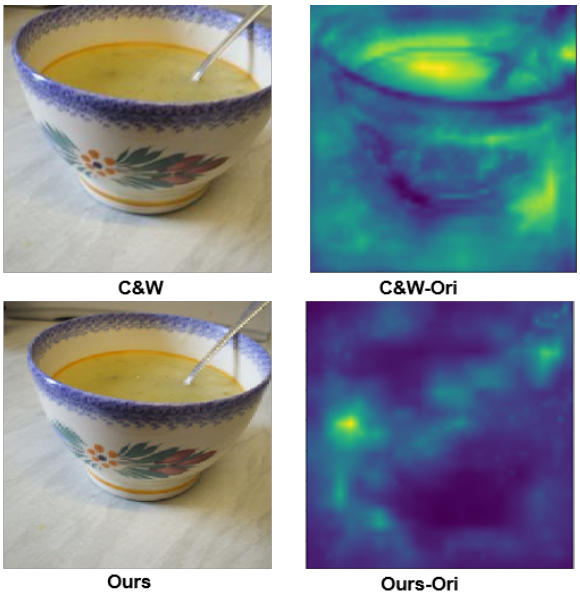

Figure 8. Comparison of perceived distance space map. As can be observed from the perceptual distance space map, DIPA generates fewer adversarial perturbations, distributed in the object edge and the background, which are less noticeable.

5. Conclusions

Our study focuses on adversarial attacks in computer vision for image sample classification tasks using a deep neural network model's attention visualization mechanism, resulting in the DIPA method. Our method is intended to generate high-quality adversarial samples, which will effectively conceal adversarial perturbations involved in attacking deep neural network models. We designed two experimental scenarios to assess the effectiveness of the DIPA. The effectiveness of utilizing the attention mechanism in combination with gradient information to search sensitive pixels in the single-pixel attack scenario is verified, and the success rate metric of the attack exceeds that of the current one-pixel attack method. Our method's ability to achieve high success rates in a short period of time in the white-box attack scenario is verified. Furthermore, the effectiveness of our method is demonstrated across various neural network models with differing architectures. Subsequently, we analyze the latent effects of adversarial perturbation using visualization and quantification techniques. We visualize the alterations induced on adversarial samples as compared to the original samples by leveraging image comparison and perceptual distance space maps. We use the Lpips perceptual quality quantification tool to assess the perceptual quality of adversarial samples and compare our results with those obtained from multiple traditional adversarial attack methods. From the comparative experimental results, the Lpips perception loss of adversarial samples generated by DIPA is lower, indicating that DIPA is more effective in hiding adversarial perturbation and adversarial samples have higher perception quality. Currently, there are many tools used to separate high-frequency information, and the methods used for image compression or frequency domain conversion can basically achieve the separation of high-frequency information. In the future, we plan to investigate the efficacy of various approaches for separating high-frequency information when generating adversarial samples. We will explore the attack effect of DIPA on other image datasets, such as CIFAR-10 and CIFAR-100 datasets, and discuss the subtle differences in DIPA attacks on different datasets. The DIPA attack proposed in this paper primarily entails the compression of image information. While the target subject information is preserved during compression, imperceptible perturbations are introduced to the high-frequency components. In our future work, we aim to identify adversarial samples by assessing the integrity of the image information. Additionally, we will attempt to reconstruct image samples using generative models to defend against such adversarial attacks.

Author Contributions: Conceptualization, J.L. and H.L.; methodology, J.L. and H.L.; software, H.L. and Y.W.; writing—original draft preparation, H.L. and P.W.; supervision, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Foundation of Inner Mongolia of China (No. 2023ZD18) and the Engineering Research Center of Ecological Big Data, Ministry of Education, the Inner Mongolia Science and Technology Plan Project (No. 2020GG0187), and the National Natural Science Foundation of China (No. 61662051).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hinton, G.E.; Osindero, S.; The, Y. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]
2. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.
3. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014; pp. 1–9.
4. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
5. Whitaker, T.A.; Simões-Franklin, C.; Newell, F.N. Vision and touch: Independent or integrated systems for the perception of texture? *Brain Res.* **2008**, *1242*, 59–72. [CrossRef] [PubMed]
6. Barreno, M.; Nelson, B.; Joseph, A.D.; Tygar, J.D. The security of machine learning. *Achine Learn.* **2010**, *81*, 121–148. [CrossRef]
7. Goodfellow, I.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 676–681.
8. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; pp. 1–11.
9. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (S&P), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
10. Cheng, S.Y.; Miao, Y.B.; Dong, Y.P.; Yang, X.; Gao, X.S.; Zhu, J. Efficient Black-box Adversarial Attacks via Bayesian Optimization Guided by a Function Prior. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; pp. 1–21.
11. Duan, R.; Chen, Y.; Niu, D.; Yang, Y.; Qin, A.K.; He, Y. AdvDrop: Adversarial attack to DNNs by dropping information. In Proceedings of the IEEE/CVF Conference on International Conference on Computer Visual (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 7486–7495.
12. Liu, D.; Su, Q.; Peng, C.; Wang, N.; Gao, X. Imperceptible Face Forgery Attack via Adversarial Semantic Mask. *arXiv* **2024**, arXiv:2406.10887.
13. Ilyas, A.; Engstrom, L.; Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019; pp. 1–13.
14. Dong, Y.; Pang, T.; Su, H.; Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4312–4321.
15. Chen, S.; He, Z.; Sun, C.; Huang, X. Universal adversarial attack on attention and the resulting dataset damagenet. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2188–2197. [CrossRef]
16. Huang, L.F.; Zhuang, W.Z.; Liao, Y.X.; Liu, N. Black-box Adversarial Attack Method Based on Evolution Strategy and Attention Mechanism. *J. Softw.* **2021**, *32*, 3512–3529.
17. Duan, M.; Qin, Y.; Deng, J.; Li, K.; Xiao, B. Dual Attention Adversarial Attacks with Limited Perturbations. *IEEE Trans. Neural Netw. Learn. Syst.* 2023, *early access*. [CrossRef] [PubMed]
18. Lin, C.; Han, S.; Zhu, J.; Li, Q.; Shen, C.; Zhang, Y.; Guan, X. Sensitive region-aware black-box adversarial attacks. *Inf. Sci.* **2023**, *637*, 118929. [CrossRef]
19. Liu, H.; Zhang, Z.H.; Xia, X.F.; Gao, T.G. A fast black box boundary attack algorithm based on geometric detection. *J. Comput. Res. Dev.* **2023**, *60*, 435–447.

20. Su, J.; Vargas, D.V.; Kouichi, S. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841. [CrossRef]
21. Papernot, N.; Mcdaniel, P.; Jha, S.; Fredrikson, M.; Swami, A.; Celik, Z.B. The limitations of deep learning in adversarial settings. In Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrucken, Germany, 21–24 March 2016; pp. 372–387.
22. Combey, T.; Loison, A.; Faucher, M.; Hajri, H. Probabilistic jacobian-based saliency maps attacks. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 558–578. [CrossRef]
23. Liu, H.L.; Liu, J. PlAA: Pixel-level Adversarial Attack on Attention for Deep Neural Network. In Proceedings of the International Conference on Artificial Neural Networks (ICANN), Bristol, UK, 6–9 September 2022; pp. 611–623.
24. Williams, P.N.; Li, K. Black-box sparse adversarial attack via multi-objective optimisation CVPR proceedings. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 12291–12301.
25. Iwana, B.K.; Kuroki, R.; Uchida, S. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In Proceedings of the International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 4176–4185.
26. Tay, Y.; Dehghani, M.; Gupta, J. Are Pretrained Convolutions Better than Pretrained Transformers? In Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL), Bangkok, Thailand, 1–6 August 2021; pp. 4349–4359.
27. Jia, D.; Wei, D.; Socher, R.; Li, L.J.; Kai, L.; Li, F.F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 398–406.
29. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer visual. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Gao, H.; Zhuang, L.; Kilian, Q.W. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
32. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
33. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef] [PubMed]
34. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Visual and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
35. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018; pp. 39–57.
36. Croce, F.; Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the IEEE Conference on International Conference on Machine Learning (ICML), Vienna, Austria, 12–18 July 2020; pp. 2206–2216.