Computer Modeling in
Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.061732

ARTICLE

# Practical Adversarial Attacks Imperceptible to Humans in Visual Recognition

**Donghyeok Park**[1], **Sumin Yeon**[2], **Hyeon Seo**[2], **Seok-Jun Buu**[2] **and Suwon Lee**[2,*]

[1]Aircraft Final Assembly Manufacturing Engineering Team, Korea Aerospace Industries, Sacheon-si, 52529, Republic of Korea
[2]Department of Computer Science and Engineering, Gyeongsang National University, Jinju-si, 52828, Republic of Korea

*Corresponding Author: Suwon Lee. Email: leesuwon@gnu.ac.kr

Received: 02 December 2024; Accepted: 03 February 2025; Published: 03 March 2025

**ABSTRACT:** Recent research on adversarial attacks has primarily focused on white-box attack techniques, with limited exploration of black-box attack methods. Furthermore, in many black-box research scenarios, it is assumed that the output label and probability distribution can be observed without imposing any constraints on the number of attack attempts. Unfortunately, this disregard for the real-world practicality of attacks, particularly their potential for human detectability, has left a gap in the research landscape. Considering these limitations, our study focuses on using a similar color attack method, assuming access only to the output label, limiting the number of attack attempts to 100, and subjecting the attacks to human perceptibility testing. Through this approach, we demonstrated the effectiveness of black box attack techniques in deceiving models and achieved a success rate of 82.68% in deceiving humans. This study emphasizes the significance of research that addresses the challenge of deceiving both humans and models, highlighting the importance of real-world applicability.

**KEYWORDS:** Adversarial attacks; image recognition; information security

## 1 Introduction

Machine learning models, particularly deep learning models, have been extensively applied in domains such as object detection in computer vision, automatic speech recognition, machine translation, and autonomous driving systems [1]. They have achieved remarkable success, often surpassing human-level performance. However, despite the transformative potential of these models, they exhibit vulnerability to adversarial attacks, which exploit weaknesses in the model to manipulate input data and produce incorrect or undesired outputs [2,3]. Such vulnerabilities pose critical risks, particularly in applications where safety and reliability are paramount, such as medical diagnostics and self-driving vehicles. For instance, altering a stop sign could mislead autonomous vehicles, which would have severe consequences [4,5].

Adversarial attacks have been a major research focus, primarily concentrating on white-box attacks, wherein the attacker assumes full access to the architecture and parameters of the model. Despite the high success rates of these models, they are less practical in real-world scenarios owing to their unrealistic assumptions. In contrast, black-box attacks, which operate without internal knowledge of the target model, are considered more practical. However, most existing black-box approaches assume access to both the output labels of the model and also the probability distribution, and they often involve thousands of attack attempts. These assumptions limit their applicability in real-world settings, wherein such extensive access and repeated attacks are infeasible.

This study addressed the aforementioned gaps by developing a novel black-box attack method, the similar color attack, which introduces several innovations to enhance realism and practicality. First, the method assumes access only to the output labels of the model, making it more aligned with real-world constraints. Second, the inefficiencies and impracticality of previous methods that rely on thousands of iterations are addressed by limiting the number of attack attempts to 100. Finally, most importantly, this study evaluated the human perceptibility of adversarial examples, a factor that has been often overlooked in existing research. While many adversarial attacks deceive machine learning models, their effectiveness against human observation remains unexamined. To this end, we surveyed with 107 participants to assess the detectability of adversarial examples generated using the proposed method.

The similar color attack method effectively deceives machine learning models by subtly modifying image colors within a predefined range, ensuring that the changes remain imperceptible to human observers. Experimental results demonstrate the high success rate of the method, achieving a 40.94% model deception rate and an 82.68% human deception rate, significantly outperforming existing techniques, such as one-pixel attacks and semantic adversarial examples. These findings highlight the potential of similar color attacks as a practical and effective adversarial technique.

By addressing both model and human vulnerabilities, this study contributes to the development of robust defense mechanisms and provides valuable insights into adversarial attack methodologies. The remainder of this paper is organized as follows. Section 2 reviews related research. Section 3 details the proposed method and its algorithm. Section 4 presents experimental results, while Section 5 discusses the implications of these findings and suggests directions for future research. Finally, Section 6 summarizes the conclusions drawn from this study.

## 2 Related Work

Adversarial attacks refer to techniques that deceive or confuse deep learning and machine learning models to improve their performance, induce misclassification, or produce unintended outputs. It is predominantly used in fields such as computer vision and natural language processing. Adversarial examples generated because of adversarial attacks serve as input data that manipulate the targeted model to produce incorrect results; thus, being employed to identify vulnerabilities and enhance stability.

Goodfellow, the creator of generative adversarial networks (GANs) [6], defined adversarial attacks and proposed methods to identify vulnerabilities and countermeasures for models. In particular, he introduced the fast gradient sign method (FGSM), which attacks models by adding small perturbations in opposite directions while searching for an optimal solution using gradient descent [3]. This suggests that the neural networks are vulnerable to small changes. FGSM is a technique that induces misclassification of models by adding small-scale noise to images. Noise is generated by applying a sign function to move the image away from the global minimum in the opposite direction. Fig. 1 shows an instance of FGSM. When the noise generated by the sign function was synthesized into an image correctly classified as a panda at a very small scale of 0.007, the image underwent a minimal change; however, the model misclassified it as a gibbon.

Subsequently, DeepFool was proposed, which iteratively queries the input vector to project it onto the decision boundary and estimates the minimum size of small changes [7], and the Carlini-Wagner (CW) attack was proposed to optimize the size of small changes and the success rate of attacks to generate the optimal adversarial examples [8]. Numerous studies on adversarial attacks are ongoing [9–11].

Various modification methods based on these attack techniques were investigated. Unlike conventional adversarial attacks that seek modifications suitable for specific images, universal adversarial perturbations operate by exploiting a network's vulnerabilities or weaknesses in a more general manner [12]. Although the

FGSM modifies input data by utilizing gradient information, the iterative fast gradient sign method (iFGSM) generates adversarial examples by iteratively accumulating small changes [13].

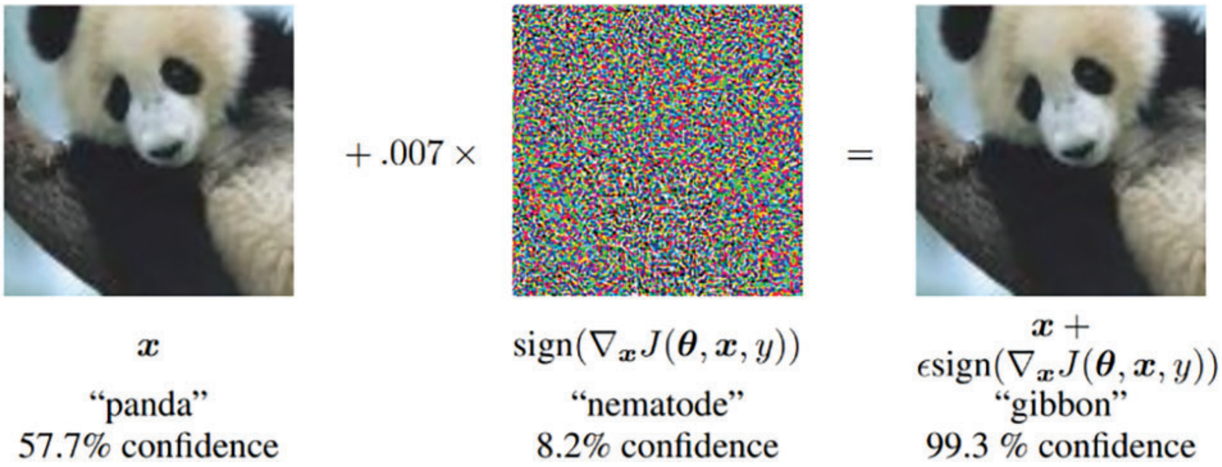
$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

**Figure 1:** FGSM's process

Research has suggested two main reasons for the existence of adversarial examples: the linearity of the learning process and statistical reasons for the high dimensionality of input images [3], and the presence of non-robust features in images [14]. It has been suggested that if the model is trained to separate the robust and no robust features of the image, the model performs normal classification.

Techniques for defending against adversarial attacks are being actively investigated. A typical defense mechanism is adversarial learning, which uses adversarial examples as training data to make a model more robust. However, this approach has limitations because it is a heuristic method and cannot defend against more powerful adversarial examples. To overcome this problem, certified adversarial robustness techniques have been proposed that mathematically guarantee that adversarial examples cannot be created within a certain boundary size. Furthermore, techniques such as randomized smoothing have been proposed, which involve training with images mixed with noise following a Gaussian distribution, ensuring that the decision boundary does not move beyond a certain boundary regardless of the direction within that boundary [15].

Adversarial attacks are categorized into white-box, black-box, and no-box attacks, based on the information available to the attacker regarding the model. White-box attacks utilize internal information such as the architecture, weights, loss functions, and training data of the target model when the attacker has full knowledge of them [2]. FGSM falls under white-box attacks [3]. White-box attacks leverage all information about the model, resulting in high attack success rates; however, they are unrealistic because it is improbable that the attacker will possess all the model information. Black-box attacks are methods in which an attacker performs an attack without knowing the internal information of the target model. The attacker provides input to the target model and observes only the results, collecting and analyzing the model's predictions to conduct the attack. There are two assumptions: one that the attacker knows both the output labels and probabilities and the other that the attacker knows only the output labels. This method is considered more realistic than white-box attacks. No-box attacks assume an environment in which the attacker cannot query a target model. Typically, an attacker creates a similar model using only a subset of the data used to train the model and attempts attacks using transferability [16]. Therefore, it can be considered the most realistic but challenging attack method to implement.

Several studies have been conducted proposing black-box attacks, which are considered more realistic than white-box attacks [17–19]. In this study, we selected the one-pixel attack [20], which is known to achieve high attack success rates while minimizing changes to the image by modifying only a single pixel, and semantic adversarial examples [21], which aim to produce adversarial examples that represent the same object as the original image, as the primary comparison targets.

The one-pixel attack is a black-box attack method in which the output labels and probabilities can be observed. It generates an initial population of 400 solutions by uniformly sampling each pixel coordinate ($x$, $y$) and pixel color (R, G, B) from the uniform and Gaussian distributions. A heuristic algorithm called differential evolution was then applied to perform 40,000 attacks per image to find the optimal attack solution. Fig. 2 shows examples of one-pixel attacks. While the original image labels were ship, horse, and deer, after the one-pixel attack, the model misclassified the images as car, frog, and airplane, with very high misclassification probabilities of 99.7%, 99.9%, and 85.3%, respectively.

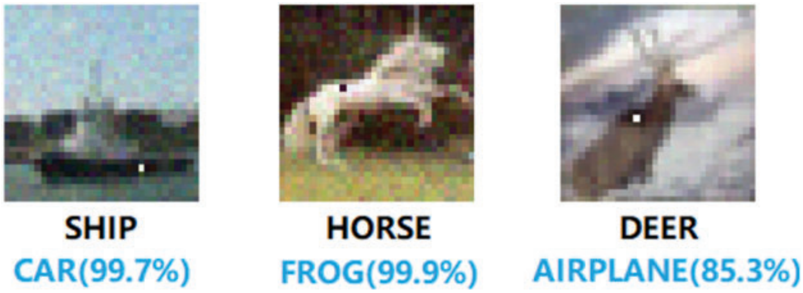

**Figure 2:** Examples of the one-pixel attack

Semantic adversarial examples are black box attack methods in which only output labels can be observed. The RGB color space of the image is transformed into the hue, saturation, and value (HSV) space, and the color and saturation components are then randomly shifted while keeping the brightness component unchanged to create adversarial examples. Fig. 3 shows examples of semantic adversarial attacks. While the original image labels were airplane, automobile, bird, cat, and deer, the model misclassified the images as dog, frog, ship, bird, and cat, respectively, after transformation into semantic adversarial examples.

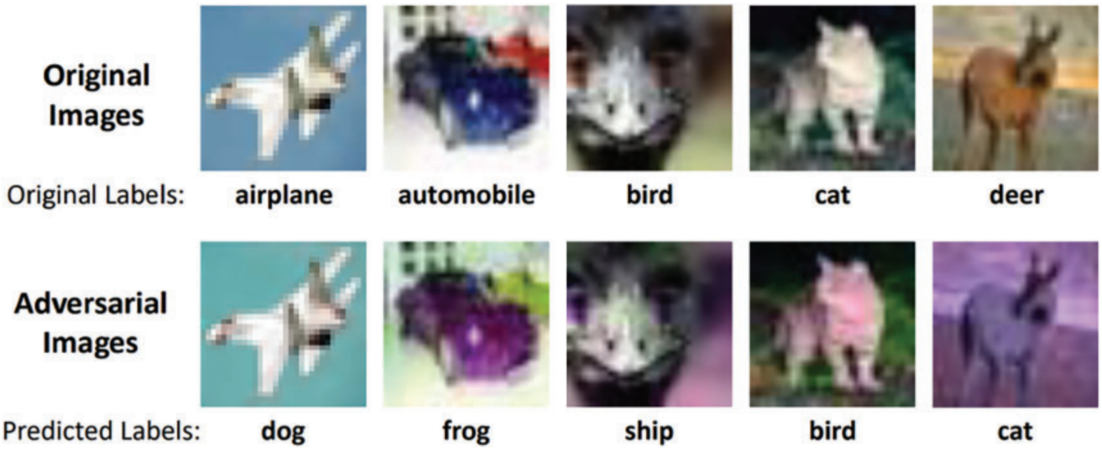

**Figure 3:** Examples of semantic adversarial examples

## 3 Proposed Method

In this section, we define similar colors that are used for a new type of attack and propose an algorithm for similar color attacks.

### *3.1 Similar Color Space*

In this study, "similar color" refers to colors that do not show significant differences in RGB channel values compared to an existing color. The range of differences in the RGB channel values was specified based on the criterion that the colors did not show significant visual differences. The range of $U_1$, representing the range within which pixels in the image change, and $U_2$, representing the maximum changed color value, are specified to change colors within a specific range. Due to these properties, it is suitable for deceiving both models and humans.

### *3.2 Similar Color Attack Algorithm*

The algorithm proposed for a realistic and feasible adversarial attack, called a similar color attack, is as follows (see Algorithm 1):

```
**Algorithm 1:** Similar color attack
**Input:** initial image $I$, image size $w$, maximum RGB value difference $U_1$, maximum RGB value difference $U_2$, maximum number of iterations $N$
**Output:** Adversarial image $I_{adv}$
Initialize $n = 1$
Pass $I_{adv}$ through the model $M$ to obtain the initial predicted label $L_{\in}$
**While** $n \leq N$ **do**
    Initialize $I_{adv} = I$
    Randomly select a pixel $P$ in $I_{adv}$
    Save each of the RGB channel values of the selected pixel: $C_{\in} = I_{adv}\,(P)$
    Find all pixel positions in $I_{adv}$ where the RGB difference from $C_{\in}$ is within $U_1$:
        $S = \{p \mid p\ is\ a\ position\ in\ I_{adv}\ and\ |I_{adv}\,(p) - C_{\in}| \leq U_1\}$
    **For** $p \in S$ **do**
        Randomly set a new RGB values $C_{new}$ such that $|C_{new} - C_{\in}| \leq U_2$
        Update the RGB values of pixel $p$ in $I_{adv}$ to $C_{new}$
        through the model $M$ to obtain the new predicted label $L_{new}$
    Pass the modified image $I_{adv}$ through the model $M$ to obtain the new predicted
        label $L_{new}$
    **if** $L_{new} \neq L_{\in}$ **then**
        **return** the adversarial image
    **end if**
    Increment $n$ by 1
**end while**
```

The algorithm begins by passing the input image $I$ into the model $M$ to obtain the initial predicted label $L_{\in}$. The adversarial image $I_{adv}$ is then initialized to be identical to $I$. A random pixel $P$ is selected from $I_{adv}$, and its RGB values are stored as $C_{\in}$. RGB values refer to the individual values of the red, green, and blue components of the RGB color channel. Next, all pixel positions in $I_{adv}$ with RGB values differing from $C_{\in}$ by at most $U_1$ are identified and collected into a set $S$. For each pixel position $p$ in $S$, a new RGB value $C_{new}$ is randomly generated such that the difference between $C_{new}$ and $C_\epsilon$ does not exceed $U_2$. The RGB values of the pixel at $p$ are then updated to $C_{new}$.

After modifying $I_{adv}(p)$, the updated image is passed into the model $M$, which outputs a new predicted label $L_{new}$. If $L_{new}$ differs from the initial label $L_\epsilon$, the algorithm terminates, and $I_{adv}$ is returned as the adversarial image. If $L_{new}$ matches $L_\epsilon$, the process is repeated, incrementing the iteration count $n$. The algorithm continues until either a successful adversarial image is generated or the maximum iteration limit $N$ is reached.

Fig. 4 shows an example of similar color attacks. A random pixel $P$ is selected from the image recognized by the model as a cat, and all pixels whose RGB values differ from $P$ by a maximum $U_1$ are found and stored in $S$. A random value $C_{new}$ is selected within a maximum range of $U_2$ from the RGB value of $P$, and the RGB values of the pixels stored in $S$ are uniformly changed to $C_{new}$. The model recognizes the modified image as that of an automobile.

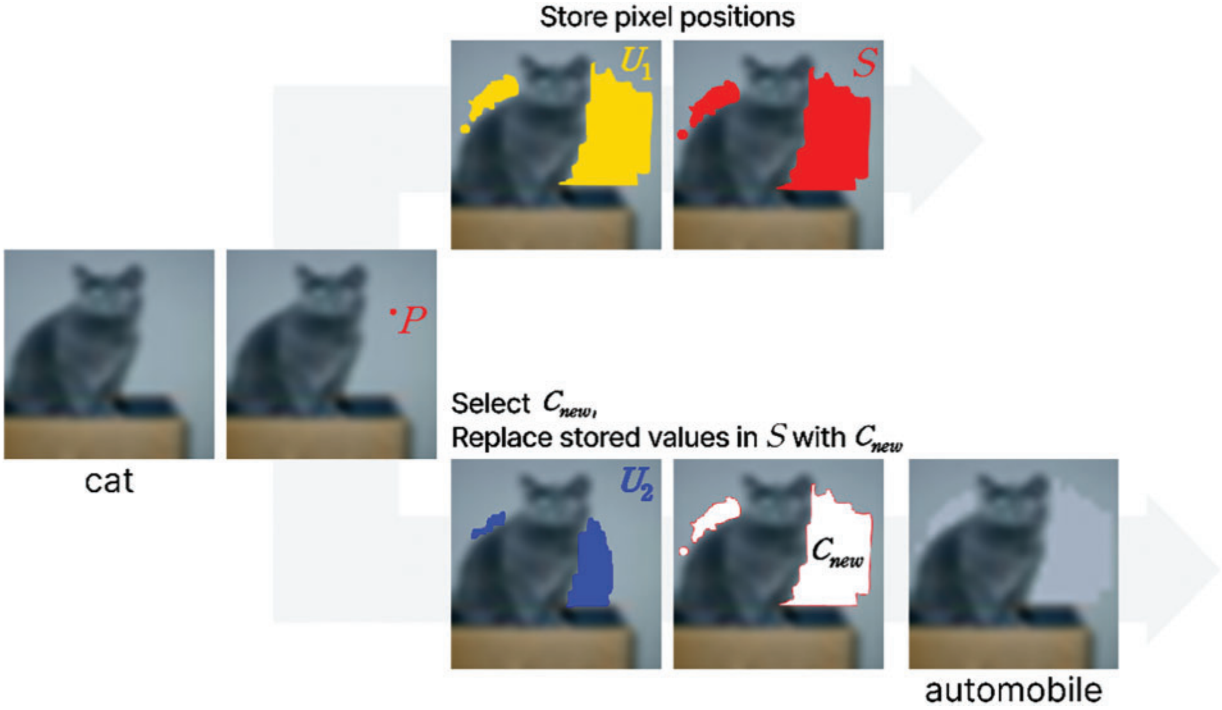

**Figure 4:** Example of similar color attack

Owing to the property of a similar color attack, even when repeating the generation of adversarial examples with the same $U_1$ and $U_2$ values, the same image is not produced. This is because, with each repetition of the attack, the criterion pixel $P$ and the new RGB value $C_{new}$ to be changed are randomly selected within a maximum range of $U_2$.

## 4 Experiments

In this section, we introduce experimental methods and results of adversarial attacks using the proposed similar color attack. We introduce the model attack experiments, followed by the human attack experiments. We chose the one-pixel attack [20] and semantic adversarial examples [21] as comparison targets.

### *4.1 Attacking Models*

This method is used to experiment with how well adversarial examples are generated by similar color attack-deceiving models. The target model is a convolutional neural network, the structure of which is shown in Fig. 5. The CIFAR-10 dataset [22] was used to train and test the proposed model. The model was trained with 50,000 images and used early stopping during the training process. The CIFAR-10 dataset was used for image classification problems, consisting of 10 classes: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks, with each image having a 32 × 32 resolution with three channels (RGB). The model test accuracy for 875 random images from the CIFAR-10 images not used for training was 80.11%.

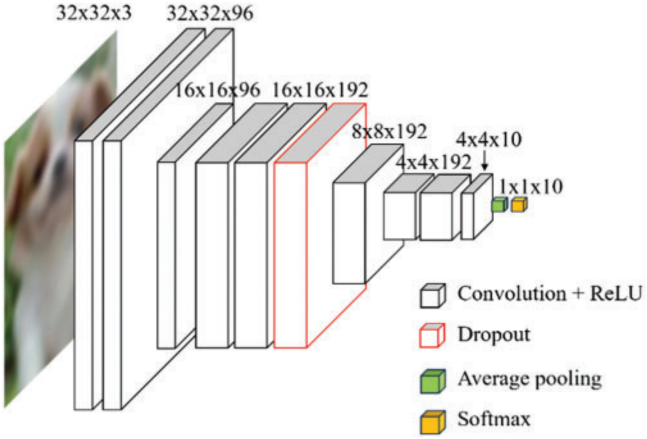

**Figure 5:** Structure of the target model

Model attack experiments were conducted as follows: For model attacks, we randomly selected 500 images from those unused in training that had the same correct label and model output label, that is, images that the model successfully classified. The maximum number of attacks was limited to 100 using a non-targeted attack method. Through multiple experiments, it was confirmed that if $U_1$, the range of pixel changes, and $U_2$, the maximum change in color values exceeds 30, significant visual distortion occurs in the original image. Therefore, within the range of 30, we selected 10, 15, 20, 25, and 30 to perform a total of 25 types of model attacks.

In this experiment, we conducted attacks using the same method and settings as those presented by the original authors of the comparison targets, one-pixel attack, and semantic adversarial example techniques. The only difference was that we adjusted the number of attacks from a maximum of 40,000 and 1000 to 100 in batches.

Fig. 6 shows examples of successful attacks using this method. "Original" means the original images, "SC" means the similar color attack, "SA" means semantic adversarial examples, and "OP" means the one-pixel attack. Whereas the original images represented a cat, airplane, frog, bird, and deer, the model perceived each adversarial example as a different image.

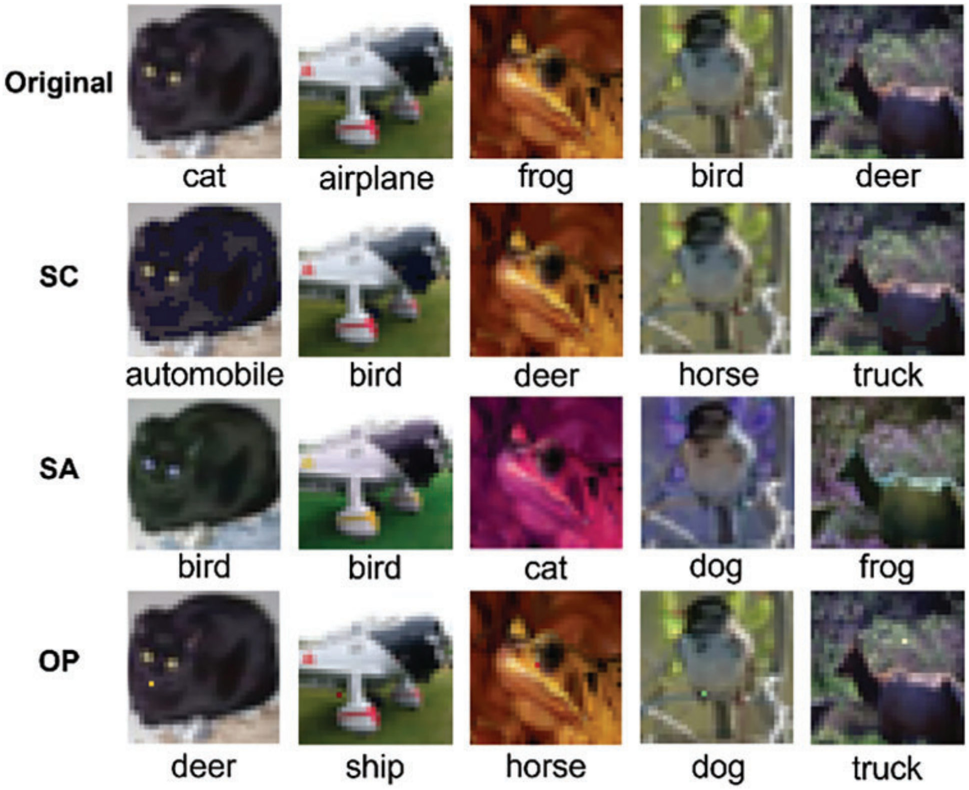

**Figure 6:** Examples of successful attacks by attack type

### *4.2 Attacking Humans*

This method was used to test how well adversarial examples generated by similar color attacks deceive humans. We randomly selected four of the 25 types of similar color attacks used in the model attack experiments described in Section 4.1 for human attack experiments. The comparison was the same as that used in the model attack experiments. The experiment was conducted using Google Forms in a survey format and involved 107 randomly selected participants who were not associated with this research. All survey participants were guaranteed anonymity and provided informed consent for the study. Participants were given a basic introduction to adversarial attacks and were then presented with each image, followed by the question: "Does the image you are currently viewing appear to be an adversarial example?".

A total of 105 images were used in the survey, including 60 images, with 15 images for each of the four types of adversarial attacks, 15 images for one-pixel attacks, 15 images for semantic adversarial examples, and 15 normal images to verify the reliability of the experimental results. None of the images overlap. Fig. 7 shows examples of the types of images used in the surveys.

The inclusion of normal images was aimed at testing the reliability of the survey. If the accuracy of classifying normal images as normal is low, it can be inferred that the respondents either lacked an understanding of adversarial attacks or responded insincerely. The accuracy of correctly classifying normal images as normal in the human-attack experiment was 88.54%. This indicated that the results of the human attack experiments were reliable. The results of the human attack experiments on images subjected to adversarial attacks are presented in the following subsection.

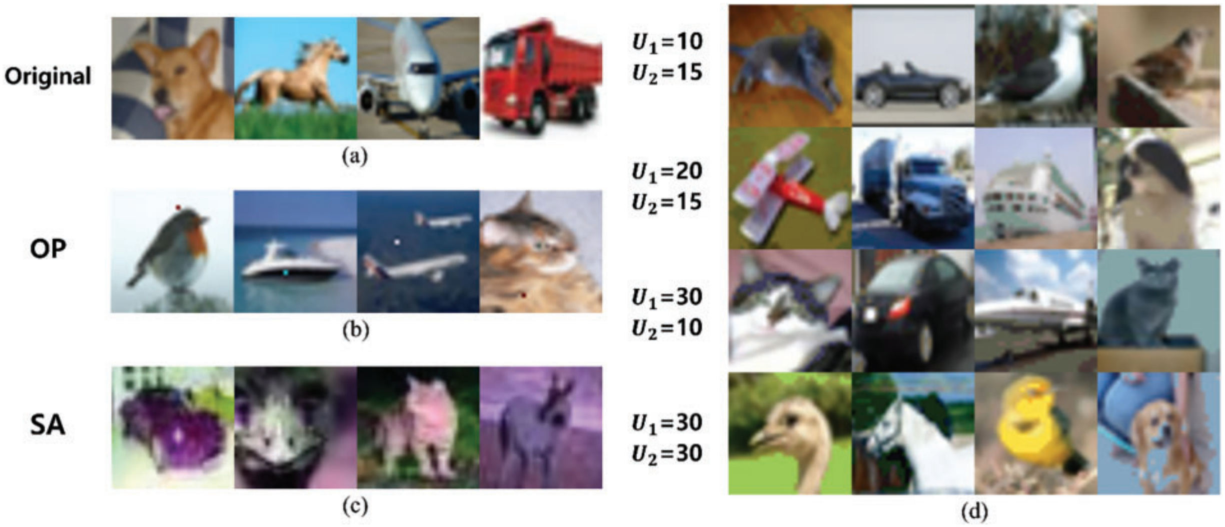

**Figure 7:** Examples of images by attack type. (a) Original; (b) One-pixel attack; (c) Semantic adversarial examples. (d) $U_1$ and $U_2$ values of similar color attack

### *4.3 Results*

The model attack success rates for similar color attacks based on the values of $U_1$ and $U_2$ are listed in Table 1.

**Table 1:** Model attack success rate by similar color attacks

| | | | $U_1$ | | |
|---|---|---|---|---|---|
| $U_2$ | **10** | **15** | **20** | **25** | **30** |
| 10 | 7.60% | 11.80% | 16.00% | 19.80% | 22.80% |
| 15 | 11.20% | 14.60% | 18.00% | 22.20% | 26.00% |
| 20 | 13.40% | 17.80% | 21.80% | 26.00% | 30.20% |
| 25 | 16.20% | 20.00% | 24.40% | 28.40% | 34.40% |
| 30 | 18.80% | 24.00% | 27.20% | 32.40% | 40.94% |

As the values of $U_1$ and $U_2$ increase, we observe an increasing trend in the model attack success rates. This is because, as adversarial examples deviate more from the original images, the model tends to classify them as different images. The upper-right area of the table shows higher attack success rates than the lower-left area, indicating that higher values of $U_1$ are more efficient in generating adversarial examples than $U_2$. In other words, when generating adversarial examples, it is more efficient to make fewer changes in many areas than to make extensive changes in a few areas. When comparing adversarial examples successfully generated by a similar color attack algorithm with the original images using the structural similarity index (SSIM) and conducting an independent sample $t$-test to determine whether there were any significant differences, no significant differences were observed between groups. In other words, images generated by a similar color-attack algorithm can be considered difficult to predict.

The model attack success rates for each attack type according to the maximum attack count $N$ are shown in Fig. 8. The $x$-axis represents the maximum attack count $N$, and the $y$-axis represents the attack success rate. For comparison, a similar color attack used $U_1$ and $U_2$ values of 30. Comparing the results based on $N$ count, it can be observed that a similar color attack has the highest attack success rate at $N = 100$ compared to other attack types.

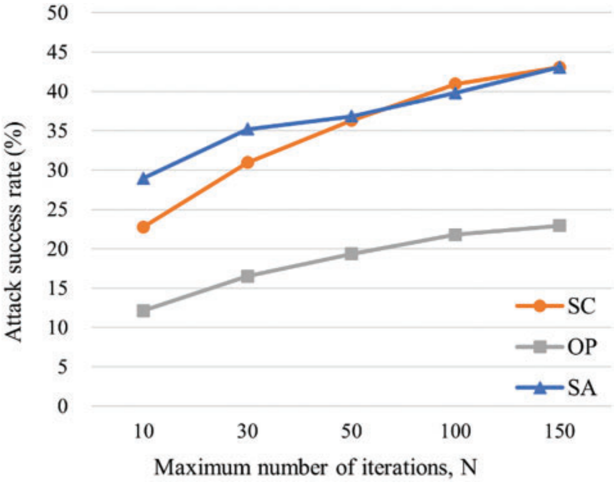

**Figure 8:** Attack success rate by number of attacks for each attack type

The experimental results for both the model and human attacks for each attack type are listed in Table 2. The attack types refer to one-pixel attacks, semantic adversarial examinations, and $U_1$ and $U_2$ values for similar color attacks. The maximum attack count $N$ was set to 100 for all cases. The numerical results indicate how effectively each attack type deceived the target, with higher numbers indicating higher attack success rates.

**Table 2:** Model and human attack success rates

| **Methods** | **Attack target** | |
|---|---|---|
| | **Model** | **Human** |
| OP | 21.80% | 57.26% |
| SA | 39.79% | 60.27% |
| SC $(U_1 = 30, U_2 = 30)$ | 40.94% | 75.08% |
| SC $(U_1 = 30, U_2 = 10)$ | 22.80% | 72.34% |
| SC $(U_1 = 20, U_2 = 15)$ | 18.00% | 72.52% |
| SC $(U_1 = 10, U_2 = 15)$ | 11.20% | 82.68% |

The experimental results showed that the success rate of human attacks with similar colors ranged from 82.68% to 72.34%. However, the success rate of human attacks is expected to increase as the values of $U_1$ and $U_2$ increase, mirroring the trend observed in model attack success rates. The experimental results did not reveal a proportional relationship between $U_1$ and $U_2$ values and human attack success rates. This suggests the importance of appropriately considering the optimal success rates of both the model and human attacks when determining $U_1$ and $U_2$ values.

Finally, a comparison of the experimental results with the comparison targets, one-pixel attack, and semantic adversarial examples shows that the model attack success rate of the one-pixel attack is 11.40%, with a human attack success rate of 57.26%, and the model attack success rate of semantic adversarial examples is 39.79%, with a human attack success rate of 60.27%. The success rate of similar color attacks was higher than that of semantic adversarial examples and one-pixel attacks. Furthermore, it is evident that the human attack success rate for a similar color attack technique significantly surpasses that of all the others. Therefore, it can be concluded that the similar color attack method proposed in this paper has proven to be highly effective not only in model attacks but also in human attacks.

## 5 Discussion

Through this study, it is evident that a similar color attack utilizing similar colors is feasible. However, this study had some limitations. The limitations of this study include the assumption of black-box attacks, allowing access to the model's output labels, and restricting attackers to 100 attempts. Therefore, several potential areas for future research are suggested:

First, research on the defense mechanisms against attack techniques that deceive both models and humans is required. Additionally, as new adversarial attack methods continue to emerge, research on their defense mechanisms is anticipated to become essential.

Second, it is essential to devise attack methods where nothing is known about the model. Some researchers have proposed no-box attack methods that utilize transferability by training new models using only a subset of the training data used during the training process. However, this assumes the availability of a portion of the data used during training and thus lacks realism. Therefore, ongoing research assuming most real-world scenarios where nothing is known about the model is necessary.

Third, although the 100 attack iterations proposed in this study are significantly fewer compared to existing studies that do not limit attack iterations, attackers need to further reduce the number of attack iterations by assuming extreme scenarios. In reality, there may be situations in which human intervention prevents attacks after detecting only a few anomalies, and objects moving rapidly, such as autonomous vehicles, may require attacks to be successful within a few seconds. Therefore, further research is required to enable successful attacks using only one attack, even in extreme cases.

Fourth, the CIFAR-10 images used in this study consist of a small number of pixels (32 × 32). Small images can easily succeed in attacks with minimal changes to a few pixels, compared to larger images. Additionally, resolution can influence not only the performance of the model but also the visual judgment of humans. Consequently, research on datasets with larger image pixel values is needed.

Future research should include comparisons not only with the attack methods examined in this study but also with various black-box attack techniques such as Natural Evolutionary Strategy (NES) attack [23], Square attack [24], and HopSkipJump attack (HSJA) [25]. These methods demonstrate high success rates in black-box scenarios and warrant experiments evaluating human detectability. Furthermore, utilizing advanced standard models like ResNet or VGG-16 could demonstrate the broader applicability of the proposed method across different architectures.

In this study, RGB channel values were used to define similar color spaces. While this approach is fast and straightforward, it has limitations in accurately calculating the differences between two colors. Future research is expected to utilize color spaces such as CIELAB to define similar color spaces with greater precision and detail.

## 6 Conclusion

In this study, we propose a similar color attack method and conduct experiments considering several realistic attack components. First, we considered black-box attacks, in which the attacker does not have knowledge of the model's architecture and can only access the output labels. Second, the number of attacks was limited to 100 iterations. Finally, to verify whether the modified image results from the attacks can evade human detection, we conducted experiments involving humans.

The experimental results demonstrate the effectiveness of the proposed method compared to existing attack techniques. Specifically, similar color attacks achieved a higher model attack success rate than one-pixel attacks and semantic adversarial examples, even under black-box constraints. Moreover, the proposed method only requires output labels rather than probability values, making it more realistic in practical attack scenarios.

Furthermore, the results of human evaluation indicate that adversarial examples generated by similar color attacks are more challenging to detect compared to those from other methods. This suggests that similar color attacks not only deceive machine learning models but also effectively evade human perception, highlighting their potential applicability in real-world adversarial settings.

In conclusion, a similar color attack method utilizing similar colors has proven to be a realistically achievable attack method, demonstrating its potential applicability in real-life scenarios.

**Acknowledgement:** None.

**Funding Statement:** This work was supported by the Research Resurgence under the Glocal University 30 Project at Gyeongsang National University in 2024.

**Author Contributions:** Study conception and design: Donghyeok Park, Suwon Lee; data collection: Sumin Yeon; analysis and interpretation of results: Hyeon Seo, Seok-Jun Buu; draft manuscript preparation: Donghyeok Park, Sumin Yeon; revision of the manuscript: Hyeon Seo, Seok-Jun Buu, Suwon Lee. All authors reviewed the results and approved the final version of the manuscript.

**Availability of Data and Materials:** The data and materials used in this study are currently part of an ongoing project and cannot be publicly released at this time. Access to the data may be considered upon reasonable request after the completion of the project.

**Ethics Approval:** This study was conducted using an anonymous online survey, collecting no personally identifiable information, including participants' names, age, or gender, to ensure full anonymity. This study did not involve any medical or psychological research on human subjects and was therefore exempt from ethical review.

**Conflicts of Interest:** The authors declare no conflicts of interest to report regarding the present study.

## References

1. Kuutti S, Bowden R, Jin Y, Barber P, Fallah S. A survey of deep learning applications to autonomous vehicle control. IEEE Trans Intell Transp Syst. 2021;22(2):712–33. doi:10.1109/TITS.2019.2962338.
2. Cha SK, Lee Y. Current research trends in attacking deep learning using adversarial examples. J KIISE. 2022;49(5):339–46. doi:10.5626/JOK.2022.49.5.339.
3. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv:14126572. 2014.
4. Duan R, Ma X, Wang Y, Bailey J, Qin AK, Yang Y. Adversarial camouflage: hiding physical-world attacks with natural styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 1000–8.

5. Thys S, Ranst WV, Goedeme T. Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2019 Jun 16–17; Long Beach, CA, USA. p. 49–55.
6. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27:1–9.
7. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2574–82.
8. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017 May 22–26; San Jose, CA, USA. p. 39–57.
9. Szegedy C. Intriguing properties of neural networks. arXiv:13126199. 2013.
10. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P); 2016 Mar 21–24; Saarbruecken, Germany. p. 372–87.
11. Zhuang J, Tang T, Ding Y, Tatikonda SC, Dvornek N, Papademetris X, et al. Adabelief optimizer: adapting stepsizes by the belief in observed gradients. Adv Neural Inf Process Syst. 2020;33:18795–806.
12. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 1765–73.
13. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. In: Artificial intelli-gence safety and security. Boca Raton, FL, USA: Chapman and Hall/CRC; 2018. p. 99–112.
14. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Mądry A. Adversarial examples are not bugs, they are features. Adv Neural Inf Process Syst. 2019;32:125–36.
15. Cohen JM, Rosenfeld E, Kolter JZ. Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA. p. 1310–20.
16. Li Q, Guo Y, Chen H. Practical no-box adversarial attacks against dnns. Adv Neural Inf Process Syst. 2020;33:12849–60.
17. Chen G, Chenb S, Fan L, Du X, Zhao Z, Song F, et al. Who is real bob? adversarial attacks on speaker recognition systems. In: 2021 IEEE Symposium on Security and Privacy (SP); 2021 May 24–27; San Francisco, CA, USA. p. 694–711.
18. Abdullah H, Rahman MS, Garcia W, Warren K, Yadav AS, Shrimpton T, et al. Hear "No evil", see "kenansville": efficient and transferable black-box attacks on speech recognition and voice identification systems. In: 2021 IEEE Symposium on Security and Privacy (SP); 2021 May 24–27; San Francisco, CA, USA. p. 712–29.
19. Zhang Y, Li Y, Liu T, Tian X. Dual-path distillation: a unified framework to improve black-box attacks. In: International Conference on Machine Learning; 2020 Jul 13–18. p. 11163–72.
20. Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput. 2019;23(5):828–41. doi:10.1109/TEVC.2019.2890858.
21. Hosseini H, Poovendran R. Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018 Jun 18–22; Salt Lake City, UT, USA. p. 1695–55.
22. Krizhevsky A. Learning multiple layers of features from tiny images. [cited 2025 Jan 10]. Available from: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.
23. Ilyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. p. 2137–46.
24. Andriushchenko M, Croce F, Flammarion N, Hein M. Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. Berlin/Heidelberg, Germany: Springer; 2020. p. 484–501.
25. Chen J, Jordan MI, Wainwright MJ. HopSkipJumpAttack: a query-efficient decision-based attack. In: 2020 IEEE Symposium on Security and Privacy (SP); 2020 May 18–21; San Francisco, CA, USA. p. 1277–94.