Hindawi
Security and Communication Networks
Volume 2021, Article ID 6637936, 11 pages
https://doi.org/10.1155/2021/6637936

Research Article

Attention-Guided Digital Adversarial Patches on Visual Detection

Dapeng Lang,[1] **Deyun Chen**,[1] **Ran Shi**,[2] **and Yongjun He**[1]

[1] *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150001, China*
[2] *School of Computer Science and Technology, Heilongjiang University, Harbin 150001, China*

Correspondence should be addressed to Deyun Chen; chendeyun@hrbust.edu.cn

Received 18 December 2020; Revised 25 January 2021; Accepted 19 March 2021; Published 8 April 2021

Academic Editor: Tom Chen

Deep learning has been widely used in the field of image classification and image recognition and achieved positive practical results. However, in recent years, a number of studies have found that the accuracy of deep learning model based on classification greatly drops when making only subtle changes to the original examples, thus realizing the attack on the deep learning model. The main methods are as follows: adjust the pixels of attack examples invisible to human eyes and induce deep learning model to make the wrong classification; by adding an adversarial patch on the detection target, guide and deceive the classification model to make it misclassification. Therefore, these methods have strong randomness and are of very limited use in practical application. Different from the previous perturbation to traffic signs, our paper proposes a method that is able to successfully hide and misclassify vehicles in complex contexts. This method takes into account the complex real scenarios and can perturb with the pictures taken by a camera and mobile phone so that the detector based on deep learning model cannot detect the vehicle or misclassification. In order to improve the robustness, the position and size of the adversarial patch are adjusted according to different detection models by introducing the attachment mechanism. Through the test of different detectors, the patch generated in the single target detection algorithm can also attack other detectors and do well in transferability. Based on the experimental part of this paper, the proposed algorithm is able to significantly lower the accuracy of the detector. Affected by the real world, such as distance, light, angles, resolution, etc., the false classification of the target is realized by reducing the confidence level and background of the target, which greatly perturbs the detection results of the target detector. In COCO Dataset 2017, it reveals that the success rate of this algorithm reaches 88.7%.

1. Introduction

The development of DNN has yielded significant results in several fields. From image processing and self-driving cars to language processing and artificial intelligence, DNN is profoundly changing the role of computers in human production and life. DNN has been able to outperform humans in many fields [1–3]. Through the analysis of DNN structure, its performance is mainly achieved by using statistical learning on a large number of data to obtain an effective representation of the input space and extract advanced features from the original sensory data. In recent years, researchers have found that DNN networks are very sensitive to disturbance resistance. Even very small data changes, such as modifying one or more pixels in an image that is invisible to the human eye, can cause DNN to misclassify or even fail to detect the target. These images with small perturbations are called adversarial examples. Initially, the researchers thought the adversarial examples were due to the nonlinearity of the deep learning model. But then Goodfellow proved through experiments that the reason was the linear behaviour of the deep neural network in high dimensional space. At the same time, it can be proved that there is no essential difference between adversarial examples and natural examples, and the examples misclassified during model testing can be used as effective adversarial examples.

In the field of image recognition [2, 3, 4–10], the CNN model learns the features of the target to be detected by learning a large number of target image models and extracting advanced features from the image. In this process, the model not only learns the distribution of features but also learns the features of the graphic background and edge. In general, the recognition accuracy will not affect the normal use of CNN. However, in the field of battlefield target recognition, vehicle identification, and security, attacks on algorithm models may have disastrous consequences. Such threats have attracted the attention of the research community and industry. Meanwhile, the method of generating adversarial examples is open and effective, such as Fast Gradient Sign Method (FGSM) [11], Projected Gradient Descent [12], DeepFool [13], etc. L-BFGS [3], as an optimization algorithm, is used in many other algorithms to generate adversarial samples. Another broad approach is to mislead the classifier by modifying a very small number of pixels in the image. There are also methods to add a patch in the image, perturbation classifier operation. This kind of patch is usually specially designed and generated. Experiments show that it is not effective to mask image features with just noise.

In practical application, there is generally no condition to modify or replace the whole image to be recognized. At the same time, the existing detector is very sensitive to the environment, angle, and clarity, and its attack effect is unstable in the experimental environment and real environment, so it is difficult to popularize in the real environment. The method of adding patch image on the image is more practical because the patch itself can be composed of various images and colours. It is not easy to attract attention, but the problem of the adversarial patch in practical application is also obvious. First of all, due to the influence of light and angle, the robustness of the patch needs to be strengthened urgently. After stretching, the patch is easy to lose its attack effect. In addition, the generation of patches depends on the algorithm structure of detection targets and detectors, so the transferability is not good. At present, the research mainly focuses on fixed small targets, such as the use of Stop signs. In this paper, electronic adversarial patches on the image are studied. In combination with the size and position of the specific target in the image, a proportional "adversarial patch" is generated, which is attached to the surface of the detection target to deceive and mislead the detector to detect the image. Figure 1 illustrates the effect.

2. Related Work

2.1. Digital Adversarial Examples. The software architecture and model structure in the white-box setting is visible to attackers, so it is easy to construct adversarial examples against classifiers. In general, through the analysis of the white-box settings, the need to be based are based on the following considerations. When attacking the photos and images of real road vehicles and cars, attackers can often obtain the detection model and algorithm structure by social engineering and reverse engineering so as to construct the adversarial examples conveniently. On the other hand, for the defence side, the analysis of attack means based on white-box setting is more conducive to the study of defence methods. In recent years, more and more researchers have applied the method of generating adversarial examples [13, 14] based on white-box setting to verify the transferability [12, 15] of black adversarial examples. By improving the white-box attack method, the migration of counter samples is better, and the black-box attack is carried out on other black-box models, which improves the effect of black-box attack. Attacks against physical objects are also on the rise [3, 11, 16–19].

Goodfellow proposed a fast gradient algorithm, which proved the existence of adversarial examples. Through the first-order approximation of the loss function [20–22], the false classification of the targets in the image was realized. In addition, the adversarial example generation method based on optimization ideas can achieve targeted attack by adding carefully chosen perturbations. The common point of these two methods is that they adopt the DAE method, namely digital adversarial examples. In the research of DAE, the current research direction mainly focuses on the detection of targets with small intraclass differences and large interclass differences, such as road STOP signs and human faces. The physical perturbations against physical targets have gradually become a research hotspot.

2.2. Object Detection. The experiment in this paper is mainly based on the target detector of YOLOv2 [6]. Its structure is based on convolution and pooling (downsampling). Finally, two fully connected layers are added. The input image is divided into $S \times S$ grids. If the coordinates of the center of the ground truth of an object fall into a grid, the grid is responsible for detecting the object. Each grid predicts B bounding boxes and their confidence scores and C category probabilities. The bounding box information (x, y, w, h) is the offset of the center position of the object relative to the grid position and the width and height, all of which are normalized. The confidence scores reflect whether the object is included and the accuracy of the position in the case of including the object, which is defined as follows:

$$\Pr(\text{Obj}) \times \text{IOU}_{\text{pred}}^{\text{truth}}, \quad \Pr(\text{Obj}) \in \{0, 1\}, \tag{1}$$

The optimization direction is to have the same prediction confidence as the ground truth IOU. The predicted conditional probability value of each grid cell is $C(\Pr(\text{Class}_i|\text{Obj}))$. In the test, each box obtains the specific category confidence by multiplying the category probability and the box confidence:

$$\Pr(\text{Class}_i|\text{Obj}) * \Pr(\text{Obj}) * \text{IOU}_{\text{pred}}^{\text{truth}} = \Pr(\text{Class}_i) \times \text{IOU}_{\text{pred}}^{\text{truth}}. \tag{2}$$

In order to train the detection network, the classification network is changed to a detection network, the last convolutional layer of the original network is removed, multiple convolutional layers (usually 1024 filters) are added, and each convolutional layer is followed by a 1×1 Convolutional layer, the number of outputs is the number required for detection. Finally, the YOLOv2 framework obtains the target class score through optimization of the cross-entropy loss function. Figure 2 shows the architecture of this process.

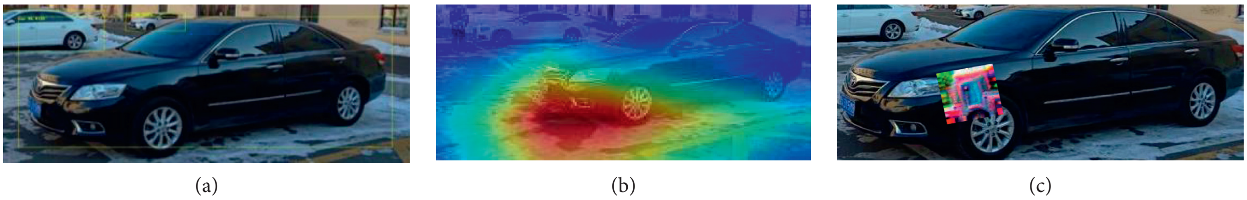


Figure 1: By detecting the key feature aggregation field in the image, the location and size of the generated adverse patch in this paper are located to achieve the purpose of the invisible detector. (a) The original car which can be detected; (b) visualize the field with high feature density; (c) the detector cannot detect the car with a patch.

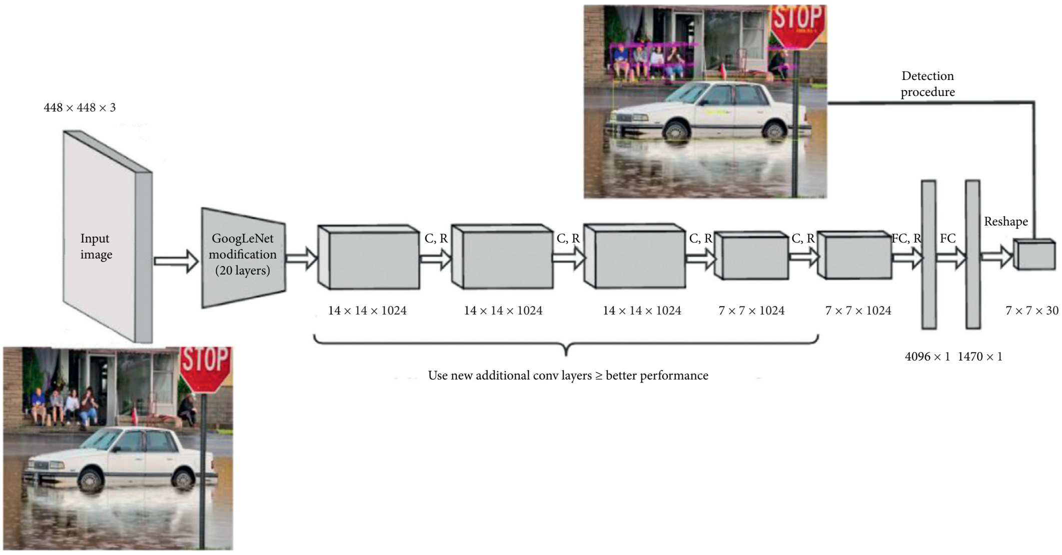


Figure 2: The typical way of detecting objects by YOLOv2.

2.3. Explainability of Attention. When adversarial patches are used as an attack method, studies have shown that the size, deformation, and illumination of the patch have a significant impact on the attack effect. However, the existing method mainly uses the random gradient descent method to generate and adjust the patch, and the effect is not stable during the application process, and the efficiency is not high. Our paper studies the formation principle of the attention mechanism [23, 24] through the method of guiding by the attention mechanism and uses the advantages of the attention mechanism from the perspective of the interpretability of the target detection algorithm. In this way, we can explore the inner workings of the deep learning framework. Attention is a mechanism used to improve the effectiveness of the Encoder + Decoder model based on RNN (LSTM or GRU). Due to the ability to distinguish attention mechanism, it is widely used in many fields such as machine translation, speech recognition, and image annotation.

In the detection process of the target detector, there is often a selective acquisition of some important parts of the observed object. Attention Mechanism can help the model assign different weights to each part of the input X, extract more critical and important information, and enable the model to make more accurate judgments without incurring greater costs for the calculation and storage. This is also the reason why the attention mechanism is so widely used. In order to fully analyse and understand the operating principle of the recognizer and provides a basis for generating adversarial patches. In this paper, the network structure is visualized in the way of class activation map visualization [25–27]. Through the heatmap, we can understand which features of the image play a key role in the image classification problem and locate the position of the object in the image.

2.4. Adversarial Patch. The adversarial patch is developed from the adversarial examples. The method of adding perturbance to the pixels of the original images makes the deep learning model like convolution neural network reduce the accuracy significantly. However, the adversarial patch does not consider whether it is perceptible to the human eyes and directly covers the "patch" on the detected target, which makes the detector misclassified or unrecognized in the detecting process [1, 18, 28, 29].

The mathematic definition would be given a patch patch, an image $x \in \text{ImagData}$, where ImagData is an image dataset, targeted classification targetClass, the location of the patch location $\in L$, the set of the transformation of images trans $\in T$, and all the space information for the rotating, scaling, etc., all of which composed of an image operator $\text{OP}(\text{patch}, x, \text{location}, \text{trans})$. The generated patch is updated iteratively by optimizing the objective function. The objective function is as follows:

$$\widehat{\text{patch}} = \overset{\text{argmax}}{\text{Patch}} E_{x\sim\text{Im}ag\text{Data},\text{trans}\sim T,\text{location}\sim L}\left[\log\ \Pr\left(\widehat{\text{patch}}|\text{OP}\left(\text{patch}, x, \text{location}, \text{trans}\right)\right)\right]. \tag{3}$$

Different from physical target hiding, the digital adversarial patch does not need to consider the dynamic environment information such as change of illumination and sudden deformation but generates the optimal adversarial structure according to the real-time snapshot. In the formula above, the image is transformed to improve the robustness of the adversarial patch so that the patch can still be effective under the operation of stretching and moving so as to complete the optimization of parameter expectation.

The adversarial patches generated in different works of the literature are shown in Figure 3 below:

3. Methods

3.1. Brief Overview of the AGAP. The goal of this paper is to generate an adversarial patch efficiently and accurately. By analysing the white-box setting of the target detection model, the algorithm generates the adversarial patch that can cheat the model detector. Based on the processing of bounding box information from the YOLO detector, the distribution of the key features of the target is analysed to form a heatmap, position, and colour based on heatmap. We attach these patches to the object that needs to be hidden in the image to hide the object. The core idea is to optimize the image at the pixel level. Based on the huge dataset, the detection score of the target detector is reduced. Our algorithm is attention-guided adversarial patch generation which is abbreviated by AGAP. The pipeline graph of our paper is in Figure 4 as follows.

3.2. Attention-Guided Feature Visualization. It is not necessary to consider whether the patch is perceptible to human eyes. The purpose of this paper is to solve the problem that the adversarial patches with a field of less than 10% are attached to images and pictures under the condition of a black-box detector, which may cause false classification or achieve unrecognized attack effect. This paper focuses on solving the two major problems in the generation of adversarial patches. The first is to construct and optimize the loss function in the process of generating the adversarial patches, which have robustness. On the other hand, we should pay attention to the location, the size, and the rotation angle of the patch. It is proved that the same patch is sensitive to location, light, size, and other factors. Even if the classifiers can be successfully attacked by patches digitally, they often do not work when they are printed or projected. Therefore, this paper studies the feature extraction process, feature meaning, and feature location of the typical CNN model used by the detector in the recognition process, forming a complementary method for patch generation. In this paper, an attention mechanism is introduced to establish the class activation graph visualization mechanism, which can visualize the key feature fields of the identified image and guide the generation and placement of adversarial patches. Figure 5 shows the process of generating a feature heatmap based on the attention mechanism. A five-layer convolution neural network is used to extract the features of the objects in the image, and a class score is used to establish class activation mapping to calculate the feature maps corresponding to different detectors.

In this paper, we use the idea of a class activation map to building a heatmap based on the attention mechanism. This scheme uses the idea of NETWORK IN NETWORK and uses global average pooling to replace the full connection layer in CNN. GAP is a special average pool layer. The reasons for adopting GAP are as follows.

Because there is no full connection layer, there is no need to adjust the size of high-definition and high-resolution images, resulting in the loss of image details. Any size image input makes it possible to support applications such as high-resolution satellite images in the future. Second, the introduction of GAP makes full use of spatial information. When the algorithm is replaced by the full connection layer, the excessive parameters of the full connection layer are deleted, which improves the robustness of the algorithm and avoids the overfitting phenomenon. Finally, in the MLPConv structure, the number of feature graphs is the same as the number of categories. After the operation of global average pooling, the results calculated by the softmax layer have a clear corresponding relationship between the feature map and the feature vector. That is, categories confidence maps are generated.

Finally, ReLU is used to calculate the output of weighted sum to ensure that the output of the algorithm only focuses on the pixel features related to classification. From the visualization of heatmap, we can see the key characteristic field of the detection target intuitively, which provides the evidence for the location, angle, and size of the patch.

3.3. Construct Loss Function. The purpose of this paper is to construct a set of adversarial patch generation algorithms which can be used to cheat the target detector. A large number of studies have proved the feasibility of adversarial patching. For objects with little change in shape, we can cheat the object detector by generating adversarial examples, but for objects with variable shapes, it is not easy to produce adversarial examples. We choose to cheat the target detector by patch, which is important because it does not need to modify the detected target directly.

At the same time, an attacker can produce different adversarial patches according to different targets in the

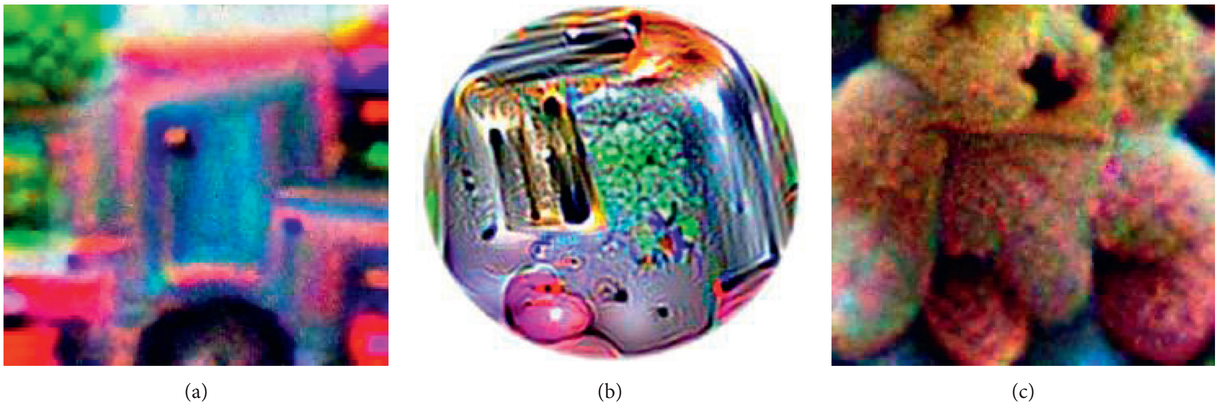


FIGURE 3: (a) Adversarial patch from our paper. (b) A toaster patch for inception-V3. (c) Adversarial patch from fooling, which looks like a bear.

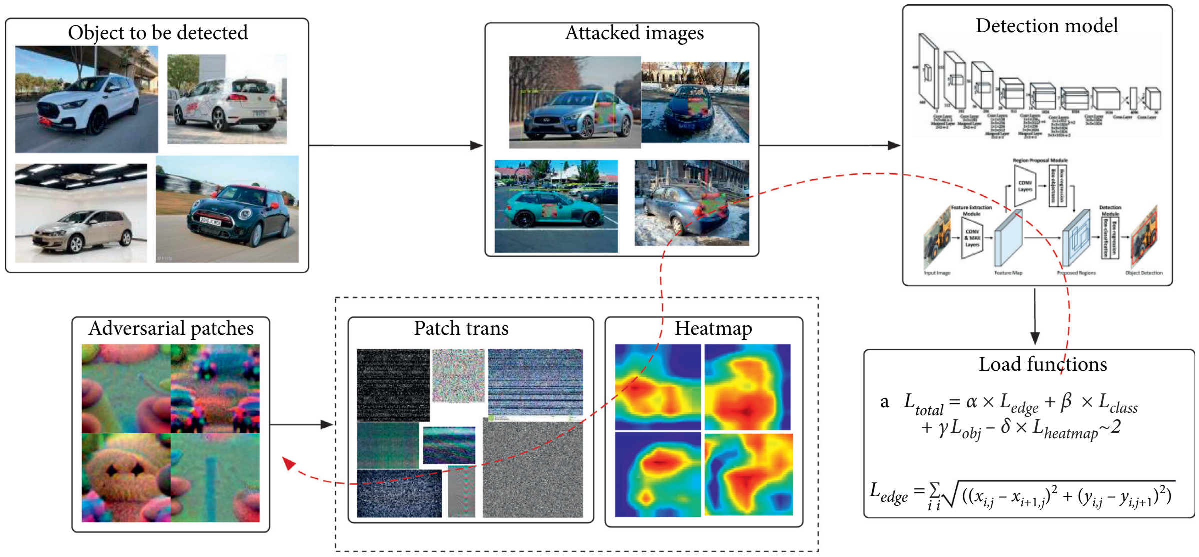

Object to be detected
Attacked images
Detection model
Adversarial patches
Patch trans
Heatmap
Load functions
a $L_{total} = \alpha \times L_{edge} + \beta \times L_{class} + \gamma L_{obj} - \delta \times L_{heatmap}{\sim}2$
$L_{edge} = \sum_{i}\sum_{i}\sqrt{((x_{i,j} - x_{i+1,j})^2 + (y_{i,j} - y_{i,j+1})^2)}$

FIGURE 4: The overview of the pipeline of our algorithm. The black arrow represents the basic process of generating an adversarial patch. The red dotted line is used to iteratively calculate the patch display form and location through backpropagation according to the loss function.

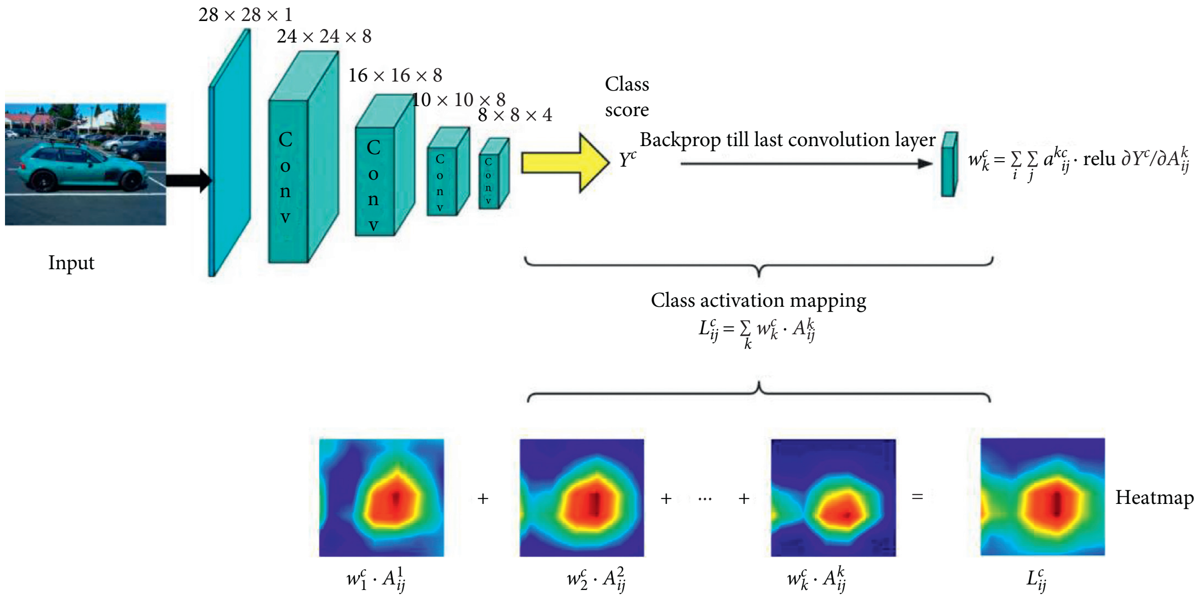


FIGURE 5: In this model, images are transformed into tensor form as the input of the visualization network. The final heatmap is obtained by feature extraction of each layer and weight. By modifying the last layer of network structure, the requirement of input image size is reduced.

process of the attack, which has a high degree of confusion. From a practical point of view, digital adversarial patches are more widely used than physical ones. They can be widely spread through the network and affect the judgment of search engines and object detectors. The loss function is optimized by iteratively updating the gradient. In the process of generating an adversarial patch, this paper carries out random translation, deformation, scaling, and other operations on the patch covered on the image to further improve the robustness of the patch and calculate its gradient loss.

Based on the principal analysis of the detector, the purpose of this paper is to greatly reduce the confidence of the detector in identifying the car. For the detector, when the confidence level of the detection is low enough, the target will weaken into a part of the background; through the optimization of classification score and object score, the attack effect is adjusted. The optimization process needs to consider the following factors:

> L_{edge}: the transition between patch edge and image should be smooth as possible. That is, the difference between patch and image needs to be optimized.

$$L_{\text{edge}} = \sum_{i,j} \sqrt{\left(\left(x_{i,j} - x_{i+1,j}\right)^2 + \left(y_{i,j} - y_{i,j+1}\right)^2\right)}. \quad (4)$$

> $L_{\text{heatmap}\sim 2}$: based on the key features of heatmap and the minimum value of L_2 norm of the patch, the randomly generated patch takes the information in heatmap as initialization information.
>
> L_{obj}: the goal of this algorithm is to make the target detector unable to recognize the vehicle in the image by constructing an adversarial patch; in this step, the L_{obj} variable needs to be minimized to reduce the score that the object detector considering that there are no objects in the detection field;
>
> L_{class}: optimize the classification score to reduce the score of the objects detected by the detector as vehicles and further control the classification score in the process of patch generation.

Based on the definition above, the total loss function is as follows:

$$L_{\text{total}} = \alpha \times L_{\text{edge}} + \beta \times L_{\text{class}} + \gamma L_{\text{obj}} - \delta \times L_{\text{heatmap}\sim 2}. \quad (5)$$

The total loss function takes the sum of the values of the three loss functions, where α, β, γ, and δ are the equation parameter, which can be set according to the expert opinion in general. By optimizing and calculating the total loss function, the optimal value of the loss function is obtained.

3.4. Data Training Strategy. In the real environment, the angle and size of the car are not fixed in the photos taken by cameras, mobile phones, and even high-resolution satellites. In the process of recognition, recognition models such as YOLOv2 and Faster-RCNN often intercept the identified target first and then establish the bounding box. Vehicle images input into the classifiers may overlap with other vehicles due to different vehicle types. Therefore, attackers must enhance their robustness through iterative optimization and multiple transformation combinations. In recent years, with the increasing research of adversarial attacks, there are more and more methods of data training and patch application in physical objects.

3.5. Experiments. We verify the effectiveness of the proposed algorithm framework. The algorithm is trained and verified based on the coco2017 data set. Coco2017 is a large image dataset released by Microsoft, which is designed for object detection, segmentation, human key point detection, semantic segmentation, and caption generation. The database collects data through extensive use of Amazon Mechanical Turk. Coco datasets now have three annotation types: object instances, object keypoints, and image captions. The coco2017 dataset includes three subsets: train (118287 images), Val (5000 images), and test (40670 images), with a total of 80 classes. Ground truth is provided for train and Val sets, but ground truth is not provided for the test set.

In our experiment, label files are used to mark the precise coordinates of each segmentation + bounding box, and the precision is two digits after the decimal point. The label of a target is as follows:

> {“segmentation”: [[392.87, 275.77, 402.24, 284.2, 382.54, 342.36, 375.99, 356.43, 372.23, 357.37, 372.23, 397.7, 383.48, 419.27, 407.87, 439.91, 427.57, 389.25, 447.26, 346.11, 447.26, 328.29, 468.84, 290.77, 472.59, 266.38], [429.44, 465.23, 453.83, 473.67, 636.73, 474.61, 636.73, 392.07, 571.07, 364.88, 546.69, 363.0]], “field”: 28458.996150000003, “iscrowd”: 0, “image_id”: 503837, “bbox”: [372.23, 266.38, 264.5, 208.23], “category_id”: 4, “id”: 151109}.

3.6. Analyse Key Feature Fields and Output Heatmap. YOLOv2 can simultaneously predict multiple bounding boxes and their class probabilities with a single convolution network. In the process of recognition, firstly, the detected image will be reduced and trimmed according to the detection target, and then multiple bounding boxes will be established for detection. After verification, although YOLO can quickly identify the target in the image, it is not accurate in locating certain objects, especially small objects. But this has little effect on the detection of vehicles in this paper.

This paper leverages the method mentioned above to calculate the feature mapping of the output of the convolution layer. In essence, these feature maps have a certain spatial correspondence with the original image. Finally, the feature map of the convolution output of the last layer is processed and redrawn to the original image to get the thermal map. The highlighted field in the heatmap is the distribution field of the key features in the network model.

Figure 6 shows the heatmap generated from the dataset based on the detection model. Among them, picture 1 is from the Internet; picture 2 below is from a mobile phone shooting.

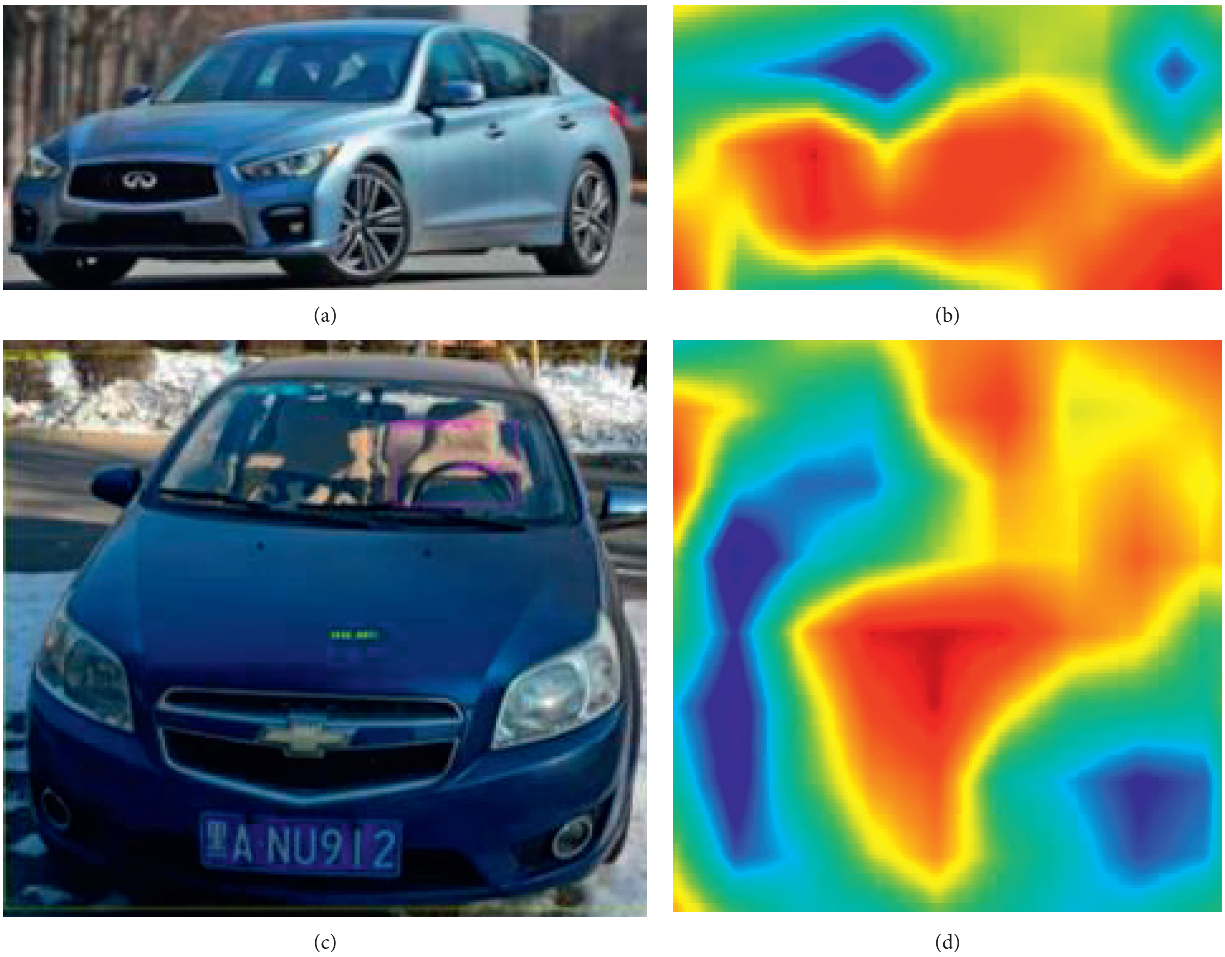


Figure 6: The two pictures above and the following two pictures, respectively, demonstrate the heatmap generated in actual operation.

3.7. Iterative Generated Adversarial Patches. To generate adversarial patches, one part of the parameters needs to be fixed, and another part of the parameters should be optimized to reduce the loss function. Because the reduction of the loss function actually affects the calculation of the object score and class score of the monitoring model, as long as the score is reduced to a certain extent, the attack effect can be achieved. If the loss function fails to converge or the target score is not well controlled in the iterative optimization process, it will lead to the misclassification of attacking missions. In this case, the attack is still successful.

Based on the analysis of the YOLO model, the target score is composed of object score and class score. The formula is as follows:

$$\text{targetScore} = \varphi \times \text{objeScore} + \phi \times \text{classScore}. \tag{6}$$

The parameters φ and ϕ represent the weight bias of the two scores adjusted according to experience in the training process. According to the result of the formula (4), with the increased number of training batches, although the human eye cannot tell the differences of patch appearances and the patch itself does not have the real physical meaning, its attack effect is improving. Figure 7 shows the adversarial patches of four batches of training with each batch training 1000 times.

3.8. Attention-Guided Patches. In this paper, we verified the generated adversarial patches by the images from different sources. Taking the picture as input, the image is processed through the detector network to identify the vehicle in the picture. The method proposed in this paper is to optimize the location, size, and angle of the adversarial patch by establishing a heatmap based on an attention mechanism. According to the introduction above, heatmap generation algorithm itself is a kind of algorithm based on recognizer network structure. Therefore, the heatmap generated for YOLOv2 is different from that generated for the Faster-RCNN model in size and layout. Therefore, the adversarial patch generated in this paper will be different in appearance and placement due to different detectors. In essence, our algorithm is a white-box-based attack algorithm. Meanwhile, the generated patch is used in Faster-RCNN as a black-box test to verify the effectiveness of this method.

As shown in Figure 8, the images in the verification experiment are mainly from four categories: images from Microsoft's coco2017 dataset; images taken by mobile phones; images downloaded from the Internet; and news pictures, in which the car is partially covered by water. The standard YOLOv2 model can accurately identify the car in the picture. According to different background

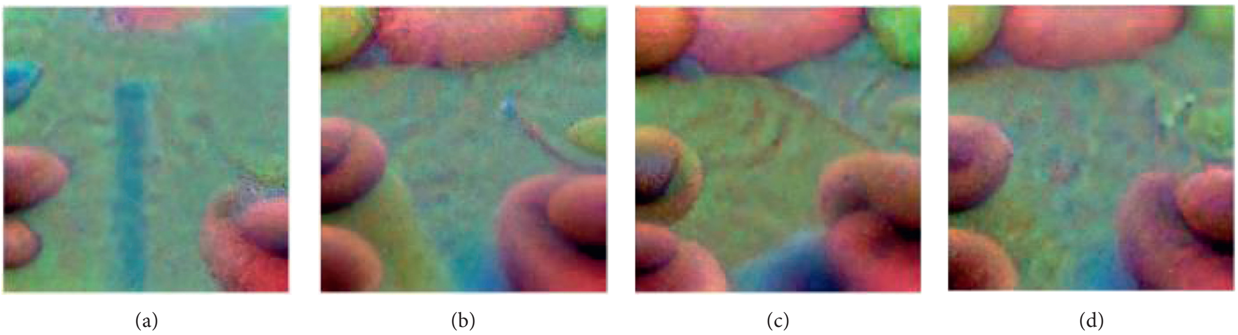


Figure 7: Four batches of adversarial patches generated by our work.

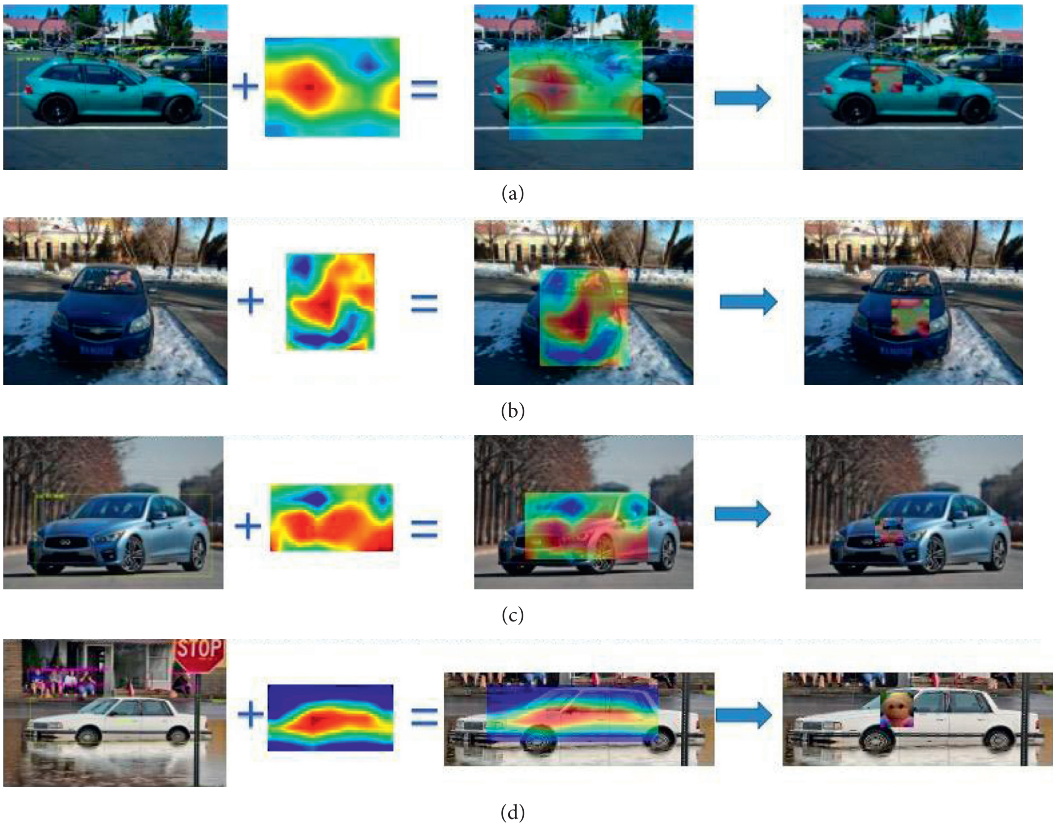
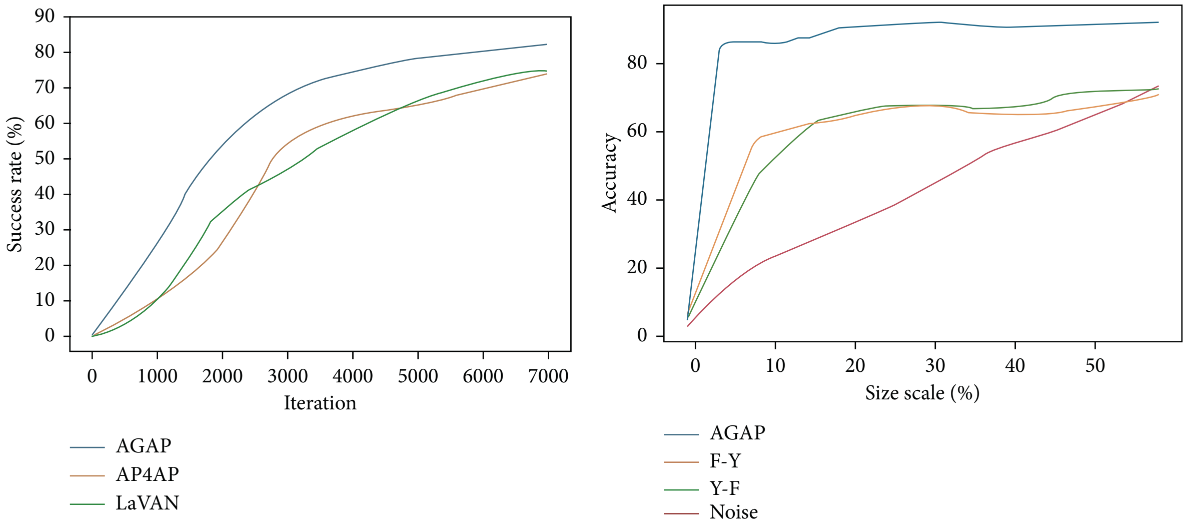


Figure 8: The flow chart generated by the attention mechanism is guided by the attention mechanism. On the images from different sources, heatmap with key features is used to limit the generation of the adversarial patch. This method can achieve the purpose of making objects invisible to a certain detector.

environments and vehicle locations, we first analyse the feature distribution extracted from the detector network and visualize it to generate heatmap. Because the detector segments the target image before detecting, the shape of the generated heatmaps is different. Then, based on the target image, we generate two kinds of adversarial patches with different appearance features. According to the location and size of the dark-red field in the heatmap, we calculate the placement position of the patches as the constraint of the loss function. According to the experimental results, it can be used as the initial location of adversarial patch near the high heat field where the heatmap and image overlap, which can generate successful patches more efficiently.

Our experimental results show that under the guidance of the attention mechanism, the interactive patch generated in this paper has strong aggressiveness and can hide the car. We have noticed that not all environments and angles can be effective for cars. In some cases, the attack effect is not ideal because the convergence speed of the loss function slows down or the feedforward gradient disappears. Then, by adjusting the value of the class score, the effect of misclassification can be produced. At the same time, replacing



(a)

(b)

Figure 9: (a) Shows the convergence speed of different patch generation algorithms under the same iteration times. In terms of hiding effect, this method is more stable and can generate available patches faster. (b) The performance of the algorithm in terms of transferability. With the increase of the adversarial patch field, the effect increases.

our generated patch with a random noise patch in the same position will not produce an adversarial effect.

3.9. Evaluations. In this paper, we propose a method of generating digital adversarial patches based on attention guidance, which can achieve the goal of hiding or misclassifying the objects in the target images. Compared with the existing research, it has obvious advantages. The existing methods of hiding targets mainly aim at scenes that are not foldable. The distance and relative position are relatively fixed, such as STOP signs. According to the existing research, the accuracy of the effective patch in the display will be greatly reduced due to the complexity of printing accuracy and illumination. Therefore, the printing of adversarial examples is also a research hotspot. Generally speaking, the main idea of the algorithm is to reduce the confidence level of the target detector. The optimization process is sensitive to the recognizer network, so it can be improved from the generation method of the patch and the structure of the detector. In terms of difficulty, because the main method is to optimize the feedback gradient, the local optimal solution may not converge, resulting in the failure to effectively hide the target.

The existing research is based on the position of the bounding box in YOLOv2 to locate the position of the patch. The initial size and position are generated randomly. The method proposed in this paper analyses the structure of the detector, extracts the key features in the process of target recognition based on the position, size, angle, and shape of these features. Attention mechanism is introduced to guide the formation of the patch. The convergence speed of the algorithm is faster when similar algorithms are used to generate the adversarial patch with the same effectiveness. The analysis is shown in Figure 9(a) below. The graph shows in 7000 training times, which shows that the algorithm proposed in this paper can achieve a stable convergence position quickly when the first 7000 times are trained, and other algorithms need more trial iterations in the process of generating patch.

About transferability testing, our paper improves the compatibility with other models by adjusting the parameters of the similar detector model. The robustness of the patch is improved by rotating and scaling the image. In this paper, the reverse patch, which is trained by the YOLO model, has been tested in the Faster-RCNN model. The transferability of patch trained in Faster-RCNN in the YOLO model is also tested, as shown in Figure 9(b). The results show that in terms of detection accuracy, although a lot of decline leads to some of the false classification results, the recognition accuracy of Faster-RCNN is still greatly reduced. The blue line in the figure is the algorithm AGAP proposed in this paper. The orange and green lines are transferability effects. Red represents the effect of noise generated patches on image hiding.

Whether the heatmap and adversarial patch based on YOLOv2 can be applied in other recognition networks also concerns us. In this paper, the same algorithm is used to calculate the heatmap of Fast-RCNN, and the effective patch of YOLOv2 is directly used to apply, but the effect is not ideal. Only about 30% of the cases are misclassified, or the image in the patch can be recognized as a target. Meanwhile, success rate of our algorithm to YOLOv2 reaches 88.7%. The comparison table of algorithm iteration and convergence is as Table 1, which includes AGAP, AP4AP, and LaVAN.

Through the research and verification of real data, we have accumulated the experience of analysing state-of-the-art detector and generating adversarial examples. It is of great significance to analyse the principle of deep neural networks in the detection model to improve the robustness and

Table 1: Comparison table of algorithm iteration and convergence.

IterTime	1000	3000	5000	7000
AGAP	27.2% (58/216)	63.7% (137/216)	80.1% (137/216)	82.9% (179/216)
AP4AP	10.4% (21/206)	45.2% (93/206)	68.2% (140/206)	81.1% (167/206)
LaVAN	15.4% (34/224)	52.1% (117/224)	76.2% (171/224)	84.5% (189/224)

antiattack of the detection model itself. In this paper, we propose an attention mechanism guided algorithm to generate an adversarial patch, which can hide and misclassify the car in an image in a controlled environment.

4. Conclusions

According to recent researches, the generation of adversarial examples is not a special game of wisdom but the inevitable result of the unique deep neural network structure. In this paper, we present an algorithm to generate adversarial ailing patches which can cheat the vehicle detector. This method takes into account the various conditions of the car in the image, including illumination, position, angle, colour, and other factors. Through the optimization of the image, the value of the loss function is continuously reduced, the confidence of the detected target is reduced, and the discrimination between the detected target and the background is reduced so as to cheat the detector.

Meanwhile, we propose a strategy to generate adversarial patches based on the attention mechanism. By extracting the key feature vectors of the car in the images, we can calculate and visualize the feature aggregation field and guide the generation and placement of adversarial patches. Finally, through a number of experiments, it is verified that in different application environments, the way of attention guidance can quickly generate adversarial patches. Experimental results show that the adversarial patches generated in this paper can attack the targeted vehicle well. After adjusting the parameters, the target can be misclassified.

The existing adversarial patch generation algorithms mainly reduce the correct classification of classifiers by randomly generating interference images. In the process of generating random adversarial patches, each pixel in the image to be detected has the same important weight, so it needs to iterate many times in the calculation process. The generated adversarial patch makes the detector unable to locate all the targets, instead of hiding an object that needs to be hidden; or when the distance angle changes, the adversarial patch might be invalid. This paper visualizes the distribution of key features through an attention mechanism. Compared with the traditional random generation algorithm, the accuracy of our algorithm is improved.

On the basis of this research, more and more researchers are paying attention to the algorithm and application of object adversarial patch generation in the reality scene. In actual production and life, it faces more and more complex problems to generate the adversarial patch for physical objects. However, once successful, the harm is also very severe. In the future, the application in this field will be able to produce a more effective, more robust, and more mobile adversarial patch for more state-of-the-art detectors, which should attract our attention.

Data Availability

All data underlying the results are available as part of the article, and no additional source data are required.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by 2020 Information Security Software Project, Network Threat Depth Analysis Software, and 2019 Industrial Internet Innovation and Development Engineering, On-Site Emergency Detection Tools in the Field of Industrial Internet Security (KY10600200021).

References

[1] Y. Wang, L. Haoran, K. Xiaohui et al., “Towards a physical-world adversarial patch for blinding object detection models,” *Information Sciences*, vol. 556, 2020.

[2] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, “This looks like that: deep learning for interpretable image recognition,” *Advances in Neural Information Processing System*, vol. 1, 2019.

[3] J. Choi, D. Chun, H. Kim, and H. J. Lee, “Gaussian YOLOv3: an accurate and fast object detector using localization uncertainty for autonomous driving,” in *Proceeding of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 502–511, IEEE, Seoul, Republic of Korea, October 2019.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proceeding of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, IEEE, Venice, Italy, October 2017.

[5] M. Xue, C. Yuan, J. Wang, and W. Liu, “DPAEG: a dependency parse-based adversarial examples generation method for intelligent Q&A robots,” *Security and Communication Networks*, vol. 2020, no. 2, 15 pages, Article ID 5890820, 2020.

[6] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in *Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6517–6525, IEEE, Honolulu, HI, USA, July 2017.

[7] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[8] A. Kwasigroch, M. Grochowski, and A. Mikolajczyk, “Neural architecture search for skin lesion classification,” *IEEE Access*, vol. 99, p. 1, 2020.

[9] N. Moritz, B. Kollmeier, and J. Anemuller, “Integration of optimized modulation filter sets into deep neural networks for automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2439–2452, 2016.

[10] Y. Fu, Y. Wei, G. Wang et al., "Self-similarity grouping: a simple unsupervised cross domain adaptation approach for person re-identification," 2019, http://arxiv.org/abs/1811.10144.

[11] K. Eykholt, I. Evtimov, E. Fernandes et al., "Physical adversarial examples for object detectors," 2018, http://arxiv.org/abs/1807.07769.

[12] L. Yujia, Z. Weiming, and Y. Nenghai, "Protecting privacy in shared photos via adversarial examples based stealth," *Security and Communication Networks*, vol. 2017, Article ID 1897438, 15 pages, 2017.

[13] J. Hayes, "On visible adversarial perturbations & digital watermarking," in *Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1678–16787, IEEE, Salt Lake City, UT, USA, June 2018.

[14] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[15] D. Lang, S. Huang, Y. Cheng et al., "A state space abstract algorithm of incremental data recognition based on model checking," *Journal of Computational Information Systems*, vol. 10, no. 4, pp. 1731–1742, 2014.

[16] H. Chang, J. Lu, F. Yu et al., "PairedCycleGAN: asymmetric style transfer for applying and removing makeup," in *Proceeding of the CVF conference on computer vision and pattern recognition (CVPR)*, IEEE, Salt Lake City, UT, USA, July 2018.

[17] J. Marniemi and M. G. Parkki, "NO need to worry about adversarial examples in object detection in autonomous vehicles," 1975, http://arxiv.org/abs/1707.03501.

[18] B. Yda, C. Xz, A. Jz et al., "Mask-guided noise restriction adversarial attacks for image classification-ScienceDirect," *Computers & Security*, vol. 560, p. 100, 2020.

[19] L. Xu, X. Wang, W. Liu, and B. Feng, "Cascaded boundary network for high-quality temporal action proposal generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3702–3713, 2020.

[20] P. Isola, J. J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, IEEE, Honolulu, HI, USA, July 2017.

[21] S. Kim, J. Jang, and O. K. Chang, "A run-to-run controller for a chemical mechanical planarization process using least squares generative adversarial networks," *Journal of Intelligent Manufacturing*, vol. 2020, pp. 1–14, 2020.

[22] L. Jiang, K. Qiao, Q. R. uoxi et al., "Cycle-consistent adversarial GAN," 2020, http://arxiv.org/abs/1904.06026.

[23] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks," in *Proceeding of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, IEEE, Lake Tahoe, NV, USA, March 2018.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[25] D. Đković, O. Golubitsky, and I. S. Kotsireas, "Some new orders of Hadamard and skew-Hadamard matrices," *Journal of Combinatorial Designs*, vol. 22, no. 6, pp. 270–277, 2014.

[26] K. Eshghi and M. Kafai, "The CRO kernel: using concomitant rank order hashes for sparse high dimensional randomized feature maps," in *Proceeding of the IEEE International Conference on Data Engineering*, IEEE, Helsinki, Finland, May 2016.

[27] F. Chen, N. Wang, J. Tang, and F. Zhu, "A feature disentangling approach for person re-identification via self-supervised data augmentation," *Applied Soft Computing*, vol. 100, no. 1, Article ID 106939, 2021.

[28] A. Liu, X. Liu, J. Fan et al., "Perceptual-sensitive GAN for generating adversarial patches," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 1028–1035, 2019.

[29] P. Wang, B. Jiao, Y. Lu et al., "Vehicle re-identification in aerial imagery: dataset and approach," in *Proceeding of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 460–469, IEEE, Seoul, Republic of Korea, September 2019.