Computers & Security 121 (2022) 102847

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

# A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions

Teng Long[a,b,*], Qi Gao[a,b], Lili Xu[b], Zhangbing Zhou[a,c]

[a] School of Information Engineering, China University of Geosciences, Beijing 100083, China
[b] Key Laboratory of Network Assessment Technology, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[c] Telecom SudParis, Institut polytechnique de Paris, Paris, France

ARTICLE INFO

*Article history:*
Received 24 April 2022
Revised 28 June 2022
Accepted 20 July 2022
Available online 22 July 2022

*Keywords:*
Deep learning
Adversarial attack
Black-box attack
White-box attack
Robustness
Visualization analysis

ABSTRACT

Deep learning has been widely applied in various fields such as computer vision, natural language processing, and data mining. Although deep learning has achieved significant success in solving complex problems, it has been shown that deep neural networks are vulnerable to adversarial attacks, resulting in models that fail to perform their tasks properly, which limits the application of deep learning in security-critical areas. In this paper, we first review some of the classical and latest representative adversarial attacks based on a reasonable taxonomy of adversarial attacks. Then, we construct a knowledge graph based on the citation relationship relying on the software VOSviewer, visualize and analyze the subject development in this field based on the information of 5923 articles from Scopus. In the end, possible research directions for the development about adversarial attacks are proposed based on the trends deduced by keywords detection analysis. All the data used for visualization are available at: https://github.com/NanyunLengmu/Adversarial-Attack-Visualization.



## 1. Introduction

In recent years, machine learning has made great progress and has been widely used in many fields. As an important branch of machine learning, deep learning is a popular research direction in the field of artificial intelligence. It has been successfully applied in classification problems (Krizhevsky et al., 2012), bioscience (Helmstaedter et al., 2013), speech recognition (Hinton et al., 2012), natural language processing (Sutskever et al., 2014), malware detection (McLaughlin et al., 2017), especially in computer vision (LeCun et al., 1998). Computer vision is the basis for many innovative key technologies that can be applied to, for example, autonomous driving (Geiger et al., 2012), intelligent industrial machines (Posada et al., 2018), and mobile applications (Howard et al., 2017), and therefore has received increasing attention which is also the focus of this paper.

Szegedy et al. (2014) first discovered the phenomena of adversarial samples in 2013, when they misclassified a deep neural network-based image classification system by adding tiny perturbations to the input samples that are undetectable to the human eye. The interference process that causes the model to misclassify is referred to as adversarial attack, and the input samples are referred to as adversarial samples in this situation. The researchers demonstrate that modern deep neural network models are highly vulnerable to adversarial attacks by small perturbations that are almost imperceptible to the human visual system, which can cause the classifier to misclassify the original image, and even worse, the attacked model will express high confidence in the output classification results. It is also proved that the same image perturbation can fool many classifiers. Adversial samples can also be used in the real world (Eykholt et al., 2018); for example, an attacker can create physical adversarial samples that prevent a traffic sign recognition system from properly recognizing warning signals or objects in self-driving vehicles from being recognized. Deep learning continues to improve and extend in different applications, but concerns about its security limit its implementation in safety-critical areas.

Many scholars have noticed the importance of the robustness of neural networks after Szegedy et al., and related research on adversarial samples has become a research hotspot in the field of deep learning. With the development of the adversarial attack field, researchers have proposed many methods to generate adversarial samples, such as increasing attack strength, improving model transfer, optimizing computing ability, etc.

The field of adversarial attacks includes many classic works, while many novel articles are being published. To systemat-

* Corresponding author.
*E-mail addresses:* longteng@cugb.edu.cn (T. Long), gaoqi1024@cugb.edu.cn (Q. Gao), xulili@iie.ac.cn (L. Xu), zbzhou@cugb.edu.cn (Z. Zhou).

https://doi.org/10.1016/j.cose.2022.102847
0167-4048/

ically understand and outline the developments in this field, there have been quite a few review articles summarizing the developments about adversarial attacks. Akhtar et al. (2021b); Akhtar and Mian (2018) provide a detailed and systematic summary of the adversarial attacks used in computer vision, who also provides a reliable summary of defensive measures. The work in Machado et al. (2023) summarizes adversarial learning for image classification from the defender's perspective and introduces the adversarial establishment and metric principles of defense. In addition, there are reviews that summarize work based on taxonomy, including work in Serban and Poll (2018), which classifies the attack strategies at that time into four categories and summarizes representative attack strategies in this way, and work in Zhou et al. (2019), who likewise reviewed the attacks at that time, and grouped them into four categories. The taxonomy-based review of attack strategies gives researchers an intuitive and systematic view of the attack development. However, the field is evolving rapidly, and many novel attack strategies have not been summarized in time by existing review work. At the same time, there is an urgent need to summarize novel and effective attack methods through rational taxonomy to supplement information about the development and frontiers of different attack strategies.

In order to better present the development of a field, visualization is an intuitive technique to achieve this goal. For trend analysis and visualization in fields, several softwares have been released to accomplish this, such as VOSviewer (Eck and Waltman, 2010), Citespace (Chen, 2006) and Bibexcel (Persson et al., 2009), etc. There are also field visualizations based on the above software (Chen et al., 2012; Meyer et al., 2014; Zha et al., 2021) in different fields of research. More practical visualization of the development of the field is made possible by the proposal of technologies such as knowledge graphs (Shaoxiong et al., 2022). Recently, Li et al. (2021b) proposed Scientific X-ray, which focuses on the scientific themes in the field of artificial intelligence by establishing a disciplinary development pipeline tree through the citation relationship data of scientific themes. With the help of Scientific X-ray, they intuitively reveal the evolutionary patterns and analyze the development potential of different themes in the field of AI. This will be an important reference for grasping research trends and showing research directions. However, the current review for adversarial attacks contains few visualization efforts for the development of the field, and there is a lack of work on building knowledge map for the development of the field. In related work, the lack of visualization has resulted in scholars not being able to more intuitively understand and grasp the situation of development in the field of adversarial attacks and to provide more diverse guidance to readers.

Based on the above-mentioned problems in the current review and the shortcomings of the related visualization work, this paper aims to introduce the current classical and newly developed attacks, and visualize and analyze the development regarding adversarial attacks. Specifically, we provide a brief introduction to the classical attacks based on the existing taxonomy of adversarial attacks, analyze and integrate the new attacks that are currently proposed. We also use dynamic network analysis techniques and visualization tool VOSviewer (Eck and Waltman, 2010) to build knowledge graph based on citation relationships for 5923 papers originating from Scopus, which are about the field of adversarial attacks, and to visualize the current works related to this field.[1] Because VOSviewer implements graphs based on co-citation and co-occurrence relationships, and supports the processing and vivid visualization of big data, enabling us to better fathom the progression of the field. In addition, research directions for the development on adversarial attacks are proposed based on the trends deduced by keywords detection analysis.

Our main contributions are as follows.

(1) We explore classical and new approaches for adversarial attacks based on taxonomies. Specifically, we refer to existing taxonomies and refine them to accommodate the latest attack strategies, including the addition of new categories. Additionally, we summarize the directions of existing research on attack strategies.

(2) We visualize and analyze the hotspots of related work about adversarial attack, based on the knowledge graphs we established, enabling a more comprehensive summary of field-related developments. In particular, we describe the process of the knowledge graph construction in detail and define the required parameters. Specifically, we analyze the literature publications, collaborations, and distribution of key articles in the field of adversarial attacks based on information from the literature and citation network, which provides a comprehensive understanding of the development concerning adversarial attack.

(3) Trends are analyzed through field keywords detection, and research directions are proposed based on the trend analysis results. Precisely, we perform keyword detection for overall field development, and in detail, we analyze the research preferences and hallmarks of different attack strategies based on keyword exploration of taxonomy. According to the keyword analysis results, we propose multifaceted field development directions for model improvement and application work in safety-critical areas, such as new scenarios, effective models, application of new technologies, etc.

The rest of this paper is as follows. In Section 2 we introduce the research work related to adversarial attacks. And we give the general structure of the main neural network models used for research and a brief definition of adversarial attacks in Section 3. After that, we analyze and summarize some of the classical attacks, and explore the latest and representative attacks in Section 4. For a more detailed review of the adversarial attack, we construct knowledge graph based on theoretical knowledge of co-citation networks for the articles with regards to adversarial attacks in Section 5. With the help of knowledge graph in Section 5, we visualize and analyze the development pertaining to adversarial attacks, ranging from the analysis of article publication, collaborative networks, to key articles in Section 6. We present keyword detection analysis and research directions for the field development in Section 7. In Section 8 we conclude the full paper.

## 2. Related work

In this section, we present related work on adversarial attacks.

Since the introduction of the adversarial sample phenomenon, there have been several works on the review of adversarial attacks on images. Serban and Poll (2018) presented a complete description of the adversarial sample phenomenon and summarized more than twenty attacks at that time by dividing the attack methods into four categories, 1) optimization-based attacks, 2) sensitive feature-based attacks, 3) geometric transformation-based attacks and 4) generative model-based attacks; Ding and Xu (2020) added functional-based attacks to the existing taxonomy. Akhtar and Mian (2018) provide a complete description of existing attacks based on attacks in classification tasks and beyond classification, and a systematic summary of defense strategies is also presented. After this, Akhtar et al. (2021a) extended the original paper in the advances based on the field of computer vision for adversarial attacks and defenses, expanding it with more recent adversarial attack defense findings. Li et al. (2022) provides experiments and summaries for typical attack and defense strategies, and provides publicly available experimental code. Machado et al. (2023) sum-

[1] All the data used for visualization are available at: https://github.com/NanyunLengmu/Adversarial-Attack-Visualization.

marizes adversarial learning for image classification from the defender's perspective, and introduces principles for building and metrics for adversarial defense. Kong et al. (2021) provides a comprehensive review of adversarial attacks from why-what-how. Qiu et al. (2019) respectively describes the corresponding adversarial attack methods from the training phase and testing phase of the adversarial attack network.

In addition, there are several review surveys on adversarial attacks in other domains besides computer vision. For the text domain, Wang et al. (2019b) classifies adversarial attacks and defense on text from the perspective of different natural language processing (NLP) tasks. For the field of adversarial on graph data, Sun et al. (2018) provides a systematic summary of existing adversarial attack and defense strategies based on graph data. For the malware identification domain, Aryal et al. (2021) providing encyclopedic introduction to adversarial attacks that are carried out against malware detection systems.

However, due to the rapid development about adversarial attacks, many novel attack methods have not been included in the existing review work, and there is a lack of analysis and overview of the current development direction of adversarial attack strategies. In addition, the current review work has not provided visual analysis and statistics of today's research progress in the field of adversarial attacks, resulting in a lack of guidance for those involved to keep abreast of relevant developments in a timely and accurate manner.

The purpose of this paper is to introduce and summarize the adversarial attack strategies about computer vision based on classification, and to summarize and analyze the latest and representative attacks. In addition, this paper will also explore and summarize the hot spots and authorities in related fields through trend analysis and visualization work based on citation network in order to serve as a guide for related researchers.

## 3. Preliminaries

In this section, we briefly introduce the general structure of Convolutional Neural Network (CNN), as CNNs are widely used in the field of computer vision and are a common model in adversarial attacks, and briefly define adversarial attacks in order to make the discussion of adversarial attacks clearer in later sections.

### 3.1. Convolutional neural network

Because Alexnet, mentioned in Krizhevsky et al. (2012) of Hinton's group in the Imagenet Image Recognition Competition in 2012, used a new deep structure of image convolution and dropout method, CNN was given renewed importance and popularity, after Yann proposed LeNet-5 in LeCun et al. (1998) in 1998, which is considered as the prototype of contemporary convolutional neural networks. A classical CNN structure has roughly five components: data input layer, convolutional layer, activation function (Rectified Linear Units layer, ReLU layer), pooling layer and output layer (fully connected layer). *Data input layer*, as the name suggests, is used to input relevant data. Before inputting the data, it is usually preprocessed to achieve better training results. *Convolutional layers* are considered as the main building blocks in a CNN model. Each convolutional layer consists of several convolutional units (filters), and the parameters of each convolutional unit are optimized by a backpropagation algorithm, which has the feature of Parameter Sharing. Convolutional layers are used for feature extraction, and more layers of convolutional layers can iteratively extract more complex features from lower-level features. *Activation function* is used to add nonlinear factors as a solution to the problem of insufficient expressiveness of linear models. The ReLU activation function was first proposed in Krizhevsky et al. (2012) and achieved excellent

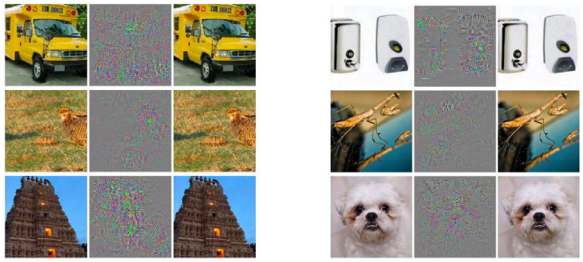



**Fig. 1.** Schematic representation of adversarial attack. Figure from Szegedy et al. (2014).

model results. *Pooling layer* slices the features into several regions and takes their maximum or average values to obtain new, smaller dimensional features to simplify the network computational complexity and extract the main features, since the model usually gets features of large dimensionality after the convolution layer. *Fully-connected layer*, which combines all local features into global features, is used to calculate the final score for each category.

In addition to AlexNet proposed by Hinton's team, CNNs also have other classical architectural forms, such as VGGNet (Simonyan and Zisserman, 2015), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), ZFNet (Zeiler and Fergus, 2014), and so on. Recently, the Meta Pseudo-Labels structure (Pham et al., 2021) based on meta-learning has gained outstanding performance and obtained over 90% accuracy in ImageNet.

### 3.2. Definition of adversarial attack

Related studies have shown that deep neural network models are vulnerable to adversarial attacks, which threaten the accuracy and security of the models. In computer vision application scenarios, images with specific perturbation noise, often called *adversarial samples*, will cause the classifier to misclassify the attacked image, which are called *target image*, while the attacked model is called *target model*. Such perturbations, called *adversarial perturbations*, are usually so extremely small that they are not detectable by the human eye, as shown in Fig. 1.

Based on the knowledge of the target model held by the attacker, adversarial attacks can usually be classified into white-box attacks, gray-box attacks, and black-box attacks. A *white-box attack* is one in which the attacker has access to all information about the target model which contains the model structure, parameters, defense strategy and control of the model input data. A *gray-box attack* is one in which the attacker is only partially informed about the target model. A *black-box attack* is one in which the attacker is unable to obtain any information about the target model and can only interact with the target model through input and output. Based on the attacker's knowledge of the target model, strategies for adversarial attacks can also be divided into various types (Serban and Poll, 2018), including gradient-Based attack (comprising optimization-based attack, etc.), transfer-based attack, score-based attack (incorporating decision-based attack, attack on attention (Chen et al., 2022a)) and geometric-transformation-based attack, etc. These categories will be described in detail in the following sections, and the latest attack strategies will be integrated and introduced.

Also, depending on the attack target, attack strategies can be classified as targeted and untargeted attacks. *Targeted attacks* have specified error classification categories, while *untargeted attacks* only require model error classification, and it is usually more difficult with targeted attacks. In addition, an attacker can also choose to use a single adversarial sample with perturbation against mul-

tiple target models, called *universal attack* (Moosavi-Dezfooli et al., 2017).

The attack success rate is an important metric to measure the quality of the adversarial attack samples, and in addition, perturbation norm are widely used to quantify the quality of the adversarial samples. Taking image data as an example, the perturbation *norm* applied in the adversarial attack includes 1)$\ell_0$, which refers to the number of modified pixel data; 2) $\ell_2$, the squared sum of the perturbed elements and then the square root, for image data, the smaller the $\ell_2$ norm indicates that the adversarial sample is harder to be recognized by human eyes; 3) $\ell_\infty$, which indicates the maximum value among the perturbed elements. Various attack methods usually make the perturbations in the adversarial sample undetectable by limiting the value of the norm.

During the attack, the attacker can use *iterative* or *one-shot* approach to find and generate the perturbation. Iteration in adversarial attacks refers to multiple computations to generate adversarial perturbations, while one shot refers to using only one computation to generate adversarial perturbations. It is generally believed that a higher number of iterations has better attack performance, while on the other hand, a higher number of iterations corresponds to more computational resource consumption.

The vast majority of existing attacks are focused on *digital attacks*, which have full access to the electronic input of the target model. In contrast, the *physical attack*, which cannot have any access to the electronic book of the target model, adds all the adversarial perturbations before generating the picture of the model input. Physical attack has important applications in scenarios such as autonomous driving security and surveillance systems.

### 3.3. Defense to adversarial attack

Adversarial defense is used to build classifiers (defense models) that are robust enough to classify correctly even when the attack image is input. There are three main approaches to building defense models (Akhtar and Mian, 2018), one is to train a more robust classifier, which is a defense from the classifier itself, commonly trained by adversarial learning. The second is to do some pre-processing of the input attack image before passing it to the classifier, with the aim of reducing the attack noise as much as possible (Liang et al., 2021). The third is to introduce additional structures to the model, such as detector, to help detect attacks better, as in Wang et al. (2019a) detected attacks by testing the sensitivity of adversarial and benign samples to random mutations.

## 4. Exploration of adversarial attacks based on taxonomy

Due to the differences in target models, perturbation methods, and test benchmarks, establishing a reasonable taxonomy for different attack methods can help scholars focus their research and reference on the required scenarios. There are also some new research applications in, for example, attention patterns that are drawing more and more interest, which are organized into some new categories in this paper to facilitate attention to developments and trend analysis in new directions. In this paper, we mainly classify the attack methods into four major categories, as shown in Fig. 2, including 1) gradient-based attack, 2) score-based attack, 3) transfer-based attack and 4) geometric-transformation-based attack. Based on the number of published articles, gradient- based attack has the highest number of publications, which proves that this strategy is more mature than others, while transfer- based attack and score- based attack have also developed well in recent years. In addition, with novel approaches such as Chen et al. (2022a) focusing on the attention to the adversarial sample, we add the taxonomy of attention-based attacks to better accommodate the summary of attacks. Since

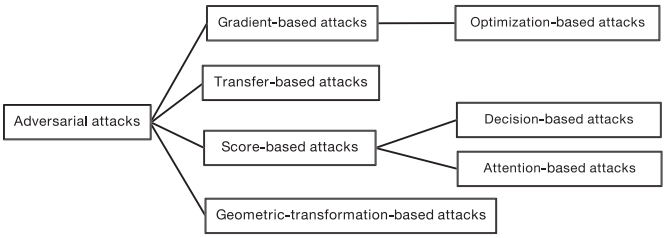



**Fig. 2.** Taxonomy of Adversarial Attacks.

Kanbak et al. (2018) first proposed geometric-transformation-based attack in 2018, the related work is still in development, but still there are many interesting methods being proposed.

In the following, we first introduce the classical attack methods in each attack category, and then introduce the more prominent and latest attack methods. The recent works summarized are presented in the tables. In Table 4.1 there are many works for obtaining more efficient and practical adversarial samples, such as more efforts are put into finding the global optimal solution of the optimization problem and attention is paid to the obtainment of smaller perturbations. To improve the query efficiency and solve the issue of insufficient training data, independently trained transfer substitute models are proposed in Table 4.2, while focusing on image features to improve the transfer success rate. Measures such as simulator models for samples are also proposed in Table 4.3 to solve the problems of inconvenient query and insufficient data for the target model. The exploration of more effective transformation approaches in images are summarized in Table 4.4, including performing fusion of different variations, etc. Note that in many cases, an attack is carried out using not only one strategy, but more often a fusion of different strategies, so the taxonomy used in this paper is for reference only.

### 4.1. Gradient-based attack

Gradient-based attack is mainly based on gradient by finding a perturbation that makes the loss value of the model larger, so that the attack sample added to that perturbation can misclassify the model. Among the current gradient-based attacks, there are mainly lines based on FGSM (Fast Gradient Sign Method) for development and improvement, and other lines for development. Since gradient-based attacks usually need to obtain information about the internal structure of the target model, the vast majority of gradient-based attacks are white-box attacks. If the gradient direction is used in the process of finding the gradient direction, it is known as optimization-based attack, which is a common way of finding the loss value.

In Table 1, we summarize and conclude some of the gradient-based attacks. It can be noticed that the exploration toward gradient-based attacks has been a popular topic for related researchers since the beginning of BFGS (Broyden-Fletcher-Goldfarb-Shanno), and it has continued till now. In recent studies, researchers have focused more on how to obtain global minimum without falling into the trap of local minimum (Wang and He, 2021), and at the same time, the searching for smaller perturbations (Zhu et al., 2021) has also become the goal of research.

In 2014, Szegedy et al. (2014) was published, a pioneering work for adversarial attacks, in which the concept of adversarial samples was first introduced. In the article, the authors point out that the input-output mapping of deep neural network learning is largely quite discreet, and that we can make the network misclassify images by applying certain imperceptible perturbations that are found by maximizing the prediction error of the network. Moreover, the specific character of these perturbations is not a random

**Table 1**
Gradient-based attacks.

| Attack | Target/Untargeted | Norm | Universal/Specific | Black/White | Iterative/one shot | Category | Year |
|---|---|---|---|---|---|---|---|
| L-BFGS (Szegedy et al., 2014) | Targeted | $\ell_\infty$ | Specific | White | Iterative | gradient | 2013 |
| FGSM (Goodfellow et al., 2015) | Untargeted | $\ell_\infty$ | Specific | White | one shot | gradient | 2015 |
| JSMA (papernot et al., 2016) | Targeted | $\ell_0$ | Specific | White | Iterative | gradient | 2015 |
| DeepFoolcite (Moosavi-Dezfooli et al., 2016) | Untargeted | $\ell_2, \ell_\infty$ | Universal, Specific | White | Iterative | gradient | 2015 |
| BIM(I-FGSM) (Kurakin et al., 2017) | Targeted | $\ell_\infty$ | Specific | White | Iterative | gradient | 2016 |
| PGD (Madry et al., 2018) | Targeted | $\ell_\infty$ | Specific | White | Iterative | gradient | 2017 |
| MI-FGSM (Dong et al., 2018) | Untargeted, Targeted | $\ell_0, \ell_2, \ell_\infty$ | Specific | White | Iterative | gradient | 2017 |
| C&W (Carlini and Wagner, 2017) | Untargeted, Targeted | $\ell_0, \ell_2, \ell_\infty$ | Specific | White | Iterative | gradient | 2017 |
| UAP (Moosavi-Dezfooli et al., 2017) | Untargeted | $\ell_2, \ell_\infty$ | Universal | White | Iterative | gradient | 2017 |
| DI2-FGSM (Xie et al., 2019) | Untargeted, Targeted | $\ell_\infty$ | Specific | White | Iterative | gradient | 2018 |
| SparseFool Attack (Modas et al., 2019) | Untargeted, Targeted | $\ell_2, \ell_\infty$ | Specific | White | Iterative | optimization | 2018 |
| ADef Attack (Alaifari et al., 2019) | Targeted | $\ell_2, \ell_\infty$ | Specific | White | Iterative | gradient | 2018 |
| EAD Attack (Chen et al., 2018) | Untargeted, Targeted | $\ell_2$ | Specific | White | Iterative | optimization | 2018 |
| EAT (Tramèr et al., 2018) | Targeted | $\ell_\infty$ | Specific | White, Black | one shot | gradient | 2018 |
| LogBarrier (Finlay et al., 2019) | Untargeted | $\ell_2, \ell_\infty$ | Specific | White | Iterative | optimization | 2019 |
| DDNA (Rony et al., 2019) | Untargeted, Targeted | $\ell_2$ | Specific | White | Iterative | optimization | 2019 |
| SI-NI-FGSM (Lin et al., 2020) | Untargeted | $\ell_\infty$ | Specific | White | Iterative | gradient, geometric-transformation | 2020 |
| VMI-FGSM (Wang et al., 2021a) | Targeted | $\ell_\infty$ | Specific | White | Iterative | gradient | 2021 |
| Homotopy-Attack (Zhu et al., 2021) | Untargeted, Targeted | $\ell_0$ | Specific | White | Iterative | optimization | 2021 |
| ALMA (Rony et al., 2021) | Untargeted, Targeted | $\ell_0, \ell_2, \ell_\infty$ | Specific | White | Iterative | gradient | 2021 |
| MGAA (Yuan et al., 2021a) | Untargeted, Targeted | $\ell_0, \ell_2, \ell_\infty$ | Universal, Specific | White, Black | Iterative | gradient | 2021 |

product of learning: the same perturbation can lead to misclassification of the same input by different networks trained on different subsets of the dataset. Also in the article, the BFGS algorithm is proposed: the problem is transformed into a convex optimization by finding the minimum loss function additive term that allows the neural network to make a misclassification.

Minimize $\|r\|_2$ subject to:

$$\begin{aligned} &f(x+r) = l \\ &x + r \in [0,1]^m \end{aligned} \tag{1}$$

Where $f(x)$ denotes the learned classification mapping function, $r$ denotes the step size of the change, and the formula expresses the search for the smallest $r$ that makes $f(x+r)$ map to the specified class $l$.

After that, Goodfellow et al. proposed FGSM (Goodfellow et al., 2015), which works in a white-box setting by finding the derivative of the model with respect to the input and then using a symbolic function to obtain its specific gradient direction, followed by multiplying by a step size, and the resulting perturbation is added to the original input to obtain the sample under the FGSM attack. The expression of the FGSM attack is as follows.

$$x' = x + \varepsilon \cdot \text{sign}\left(\nabla_x J(x, y)\right) \tag{2}$$

In describing the principle of their operation, the authors explain that the effects caused by adversarial perturbations are amplified in deep neural networks, especially in linear models, while current neural network constructions usually tend to use linear activation functions like Relu, making the network as a whole converge to linearity. In addition, they propose that the larger the dimensionality of the model input, the more vulnerable the model is to attack. The FGSM algorithm is simple and effective, and he makes the target model produce 89.4% misclassification on the MNIST dataset, which plays a very important role in the field of image attacks, and many subsequent studies have been carried out based on this algorithm.

Since the FGSM algorithm involves only a single gradient update and sometimes a single update is not enough for a successful attack, the BIM (Basic Iterative Method, also known as iterative FGSM) Kurakin et al. (2017) is proposed, which gets the attack samples by continuously iterating the FGSM algorithm to obtain attack samples for better attack effect. The attack expression of BIM is as follows.

$$\begin{aligned} &\boldsymbol{X}_0^{adv} = \boldsymbol{X} \\ &\boldsymbol{X}_{N+1}^{adv} = \text{Clip}_{X,\varepsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha\ \text{sign}\left(\nabla_X J\left(\boldsymbol{X}_N^{adv}, y_{true}\right)\right)\right\} \end{aligned} \tag{3}$$

In this attack, each time the individual pixel grows (or decreases) by $\alpha$ based on the adversarial sample from the previous step, and then it is clipped to ensure that each pixel of the new sample is within the $\varepsilon$ critical region of each pixel of $X$, in order to make the adversarial sample found with the change of each pixel less than $\varepsilon$. BIM is considered one of the most powerful attacks due to the multiple searches for effective perturbations, but it is considered computationally expensive. Later, Madry et al. (2018) proposed PGD (Projected Gradient Descent), which is a variant of BIM. Compared with the BIM algorithm, it initializes with uniform random noise, increases the number of iteration rounds, and proposes to use projection against gradients instead of clip operation on gradients in BIM. After experimental validation, PGD is considered to be probably the most powerful first-order attack.

To solve the uncertainty problem of $\varepsilon$ in FGSM, Moosavi-Dezfooli et al. (2016) proposed DeepFool, which is based on hyperplane classification, and for the first time proposed to obtain the minimal perturbation against the model by measuring the closest distance between the sample and the decision boundary. Deepfool

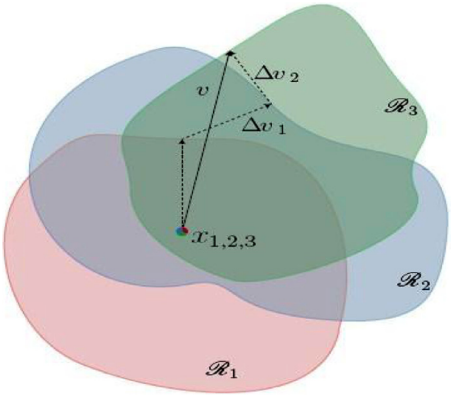



**Fig. 3.** Schematic of the process of finding Universal adversarial perturbations. Figure from Moosavi-Dezfooli et al. (2017).

can attack both binary and multi classification, linear and nonlinear classifiers. Since it generates (approximate) minimum perturbation, DeepFool can be used to estimate the robustness of the model.

Unlike FGSM, which utilizes the gradient information of the loss function of the model output, JSMA (papernot et al., 2016) (Jacobian-based Saliency Map Attacks) introduces the concept of Saliency Map, which mainly uses the output category probability information of the model to back propagate the corresponding gradient information, and finds the corresponding perturbation by observing the effect of the input perturbation on the output result. If some features are found to correspond to a specific output in the classifier, the algorithm will enhance or weaken these features in the input samples in a greedy way to make the classifier produce the specified output. The authors propose "forward derivative" to obtain the direction of the gradient of the predicted values of the target class tokens. In this attack, for the first time, the perturbation norm is controlled at $\ell_0$, i.e., the number of pixels is modified as little as possible to obtain a better attack.

Meanwhile, unlike the gradient-based approach of FGSM that restricts the perturbation size in each step, Carlini and Wagner (2017) (C&W) proposed three different regularized attack methods ($\ell_0, \ell_2, \ell_\infty$), which introduce the problem of generating adversarial samples into the problem of finding the minimum perturbation problem. This method is extremely slow due to the need to optimize some of the parameters in this algorithm, and this method does not have black-box transferability, but this method is a very strong white-box attack method and is resistant to defensive distillation.

In addition, Moosavi also proposed an interesting and practical attack UAP (Universal adversarial perturbations) (Moosavi-Dezfooli et al., 2017), where the attacker only needs to add perturbations under this universal algorithm to all samples of the same distribution to achieve adversarial sample construction, solving the previous the problem of customizing the perturbations of the algorithm. As shown in Fig. 3, the algorithm iteratively derives the launch perturbation vector $\Delta v$ for each training sample in turn, i.e., it eventually generates a perturbation that can jump out of the decision boundary of the set of all training samples.

*Variance tuning:* Wang and He (2021) proposed a new method called variance tuning to enhance the class of iterative gradient-based attack methods and improve their attack transferability. Specifically, instead of directly using the current gradient for momentum accumulation in each iteration of the gradient computation, the current gradient is further adjusted by considering the gradient changes from the previous iteration, thus destabilizing the update direction and getting rid of the local optimum. The key idea is to reduce the gradient change at each iteration, so as to stabilize the update direction and get rid of the local optimum solution during the search process. This method can achieve an average attack success rate of 90.1% in the face of nine defense methods with input transformations and a multi-model setup, which improves the best attack of the moment by 85.1%.

*Homotopy-attack:* Sparse adversarial attacks can deceive deep neural networks by perturbing only a few pixels. Compared to pixel-wise, highly sparse adversarial attacks are more dangerous because are less detectable. Zhu et al. (2021) jointly tackle the sparsity and the perturbation bound in one by using the homotopy algorithm unified framework. The method exploits the properties of different regions to impose different degrees of infinite norm perturbation upper bound, where the computation of this bound relies on the pixel saturation levels of different axes to minimize the $\ell_0$ distance between the minimized adversarial samples and the clean samples. Experiments show that the method can produce very sparse adversarial perturbations while maintaining a relatively low perturbation strength compared to state-of-the-art methods.

### 4.2. Transfer-based attack

Transfer-based attack does not rely on information about the target model, but requires information about the training data. Szegedy et al. (2014) first proposed that an adversarial sample generated against one model can be transferred to another model and can cause an effective attack. Lu et al. (2017) demonstrated that if attack samples are created on a set of alternative models, the success rate of the attacked model can reach 100% in some cases. transfer-based attack is a attack between black-box and white-box attack.

In Table 2, we summarize and conclude some of the transfer-based attacks. Better transferability is an important goal in the research of Transfer-based attacks, and in recent studies related people have accomplished this goal by adversarial generative networks (GANs) (Mingyi et al., 2020), feature extraction (Wang et al., 2021b), etc. to accomplish this goal. So far, transfer-based attacks all require iterative generation to get a viable attack.

Papernot et al. (2017) first proposed a black-box-based attack approach by generating substitute models to simulate the decision boundaries of the approximated attacked model, and generating adversarial samples based on the current alternative model, and these adversarial samples are eventually used to attack the original target model. During the training process, the Jacobi matrix is used to efficiently utilize the query results in order to reduce the number of queries for the target model. This method invalidates the gradient-mask defense strategy because it does not require gradient information. Later, Liu et al. (2017) introduced the idea of ensemble in this method, i.e., selecting multiple models simultaneously and combining their loss values to generate the corresponding adversarial samples. This method takes into account the similarity of decision boundaries among different models, and thus achieves the goal of migrating a large range of adversarial samples among different models for the first time. In addition, Huang and Zhang (2020) proposed TREMBA combining transfer-based and scored-based attack ideas. The method firstly generates a preliminary adversarial sample in the white-box attack by the substitute model, and then uses this preliminary adversarial sample as the search starting point, continues to query using the score-based attack method, and finally iterates the final adversarial sample with good migration. This method effectively reduces the number of queries while improving the success rate of black-box attacks.

*DaST:* Mingyi et al. (2020) proposed DaST (Data-free Substitute Training), a method that does not require data to train a substitute

**Table 2**
Transfer-based attacks.

| Attack | Target/Untargeted | Norm | Universal/Specific | Black/White | Category | Year |
|---|---|---|---|---|---|---|
| SafetyNet(Lu et al., 2017) | Targeted | $\ell_2, \ell_\infty$ | Specific | Black | transfer | 2016 |
| substitute(Papernot et al., 2017) | Targeted | $\ell_2$ | Specific | Black | transfer | 2017 |
| Ensemble(Liu et al., 2017) | Untargeted, Targeted | $\ell_2, \ell_\infty$ | Specific | Black | transfer | 2017 |
| ILA(Huang et al., 2019) | Untargeted, Targeted | $\ell_\infty$ | Universal, Specific | Black | transfer | 2019 |
| P-RGF(Cheng et al., 2019) | Untargeted, Targeted | $\ell_2, \ell_\infty$ | Specific | Black | transfer | 2019 |
| TREMBA(Huang and Zhang, 2020) | Targeted | $\ell_\infty$ | Specific | Black | transfer, score | 2020 |
| ATA(Wu et al., 2020) | Targeted | $\ell_0, \ell_\infty$ | Specific | Black, White | transfer, attention | 2020 |
| DaST(Mingyi et al., 2020) | Untargeted, Targeted | $\ell_\infty$ | Specific | Black | transfer | 2020 |
| FIA(Wang et al., 2021b) | Targeted | $\ell_2$ | Specific | Black | transfer, gradient | 2021 |

model to achieve adversarial attacks. It uses generative adversarial networks (GANs) to generate synthetic samples to train the substitute model, whose synthetic samples are labeled from the target model. At the same time, to solve the problem that traditional GANs may generate extremely uneven distribution of samples if no real data is available, the authors design a multi-branch architecture and a loss function controlling the labels for the generative model to solve the problem of uneven distribution of synthetic samples. The adversarial samples under this method have excellent transferability. The FE-DaST proposed by Yu and Sun (2022) used a single branch generator to obtain better model similarity and attack success based on information entropy loss, by building simpler models. Furthermore, FE-DaST is able to be used on a larger number of datasets compared to DaST.

*FIA:* In the attacks proposed in the past, the model treats the points in the picture equally without differentiation and learns many noise features that lack transferability, which easily leads to local optimality. Wang et al. (2021b) proposed Feature Importance-aware Attack, which uses gradients to represent the importance of features and optimizes the weighted feature mapping by suppressing positive (important) features and promoting negative (trivial) features to make model decisions wrong, resulting in higher transferable adversarial samples. Experiments show that FIA has outstanding black-box attack effectiveness.

### 4.3. Score-based attack

Score-based attack is black-box attack, which relies only on prediction scores (e.g., category probability or logarithm) for the prediction of the gradient. In many cases, the attacker does not have access to the prediction score of the target model, but only to the classification results of the corresponding samples, and relies only on the decision boundary to perform the attack, then this type of attack is called decision-based, which is considered to be more practical. Also, when the classification error is achieved by shifting the attention to the target label, this method is known as attention-based attack.

In Table 3, we summarize and conclude some of the score-based attacks. The latest research usually focuses on attacks with less query complexity and higher attack transferability by means of meta-learning Du et al. (2019), building simulators Ma et al. (2021), etc. Since only queries on the target model are needed to obtain the bound information, all score-based attacks are black-box. Also effective attacks require multiple queries, so the generation of perturbations is iterative.

Chen et al. (2017) proposed ZOO (Zeroth Order Optimization), which started the research trend of score-based attack. The attack finds perturbation samples by obtaining the probability of each label of that sample under the input and target model. The attack first estimates a gradient value and then uses an optimization method such as Newton's method or adam to obtain the optimal gradient, which is superimposed on the image and then input to the model. The attack stops after a successful attack, otherwise it continues iteratively to estimate the gradient.

$$\begin{aligned} &\text{minimize}_{\mathrm{x}} \|\mathrm{x}-\mathrm{x}_0\|_2^2 + c \cdot f(\mathrm{x}, t) \\ &\textit{subject to } x \in [0,1]^p \end{aligned} \tag{4}$$

Where $\mathrm{x}_0$ denotes the original image, $x$ denotes the modified image, $t$ denotes the redirected label, and $f(x,t)$ denotes the loss function (or confidence) of $x$ classified as $t$. This transforms the adversarial attack problem into an optimization problem that minimizes the sum of these two. Then Su et al. (2019) proposed One Pixel Attack, which has high picture utility because it only needs to change fewer points or one pixel point to obtain a better attack effect. To improve the efficiency of finding the attacked pixel points, the finding strategy of differential evolution is introduced. Also the attack only needs to obtain the label probability of the black box without using the internal parameters of the network, and it can attack the models that are non-differentiable or the gradient is difficult to calculate, which makes the attack strategy have better practicality.

*Query-efficient meta attack:* Du et al. (2019) uses meta-learning based on autoencoder structure to approximate the gradient and use reptile meta-learning training method for training. By training the mata attacker and incorporating it into the optimization process, the method can significantly reduce the number of queries required without reducing the success rate and distortion of the attack.

*Attack on attention:* The AoA (Attack on Attention) approach proposed by Chen et al. (2022a) is an improvement of the score-based method. Unlike the score-based method, AoA attacks the attention heat map, a common semantic feature among networks, to shift the attention from the original class (non-target class) to close to the target class (target), thus making the classifier work incorrectly. The method achieves the best black-box attack migration success rate so far in image classification neural networks. The authors also constructed an adversarial test set DAmageNet based on AoA to help researchers perform relevant robustness tests.

*Simulator attack:* Ma et al. (2021) trains a simulator where MSE (Mean Squared Error) loss functions based on knowledge distillation are applied to internal and external updates in meta-learning to learn the outputs of many different network models, thus allowing simulate the output of any unknown model. Once trained, the simulator requires only a small amount of query data for fine-tuning to accurately simulate the output of the unknown network, thus making a large number of queries to be transferred to the simulator, effectively reducing the query complexity of the target model in the attack.

### 4.4. Geometric-transformation-based attack

Geometric-transformation-based attacks generate adversarial samples by performing geometry-based operations (Wang et al., 2021a) (angles, scaling, shifts, etc.), color-based operations (Chen et al., 2022b) (brightness, color, contrast, etc.), or synthetic transformations (Admix-based) on the samples, where

**Table 3**
Score-based attacks.

| Attack | Target/Untargeted | Norm | Universal/Specific | Category | Year |
|---|---|---|---|---|---|
| ZOO(Chen et al., 2017) | Untargeted, Targeted | $\ell_2$ | Specific | transfer, score | 2017 |
| UPSET(Sarkar et al., 2017) | targeted | $\ell_2$ | Universal | score | 2017 |
| ANGRI(Sarkar et al., 2017) | targeted | $\ell_2$ | Specific | score | 2017 |
| Boundary Attack(Brendel et al., 2018) | Untargeted, Targeted | $\ell_2$ | Specific | decision | 2018 |
| qFool(Liu et al., 2019) | Untargeted, Targeted | $\ell_2$, $\ell_\infty$ | Specific | decision | 2018 |
| One-Pixel Attack(Su et al., 2019) | Untargeted, Targeted | $\ell_0$ | Specific | decision | 2019 |
| AutoZOOM(Tu et al., 2019) | Untargeted, Targeted | $\ell_2$, $\ell_\infty$ | Specific | score | 2019 |
| CornerSearch(Croce and Hein, 2019) | Untargeted, Targeted | $\ell_0$ | Specific | score | 2019 |
| Trust Region(Yao et al., 2019) | targeted | $\ell_2$, $\ell_\infty$ | Specific | decision | 2019 |
| PCA Attack(Wang et al., 2019c) | Untargeted, Targeted | $\ell_2$, $\ell_\infty$ | Specific | decision | 2019 |
| Avolutionary Attack(Dong et al., 2019b) | Untargeted, Targeted | $\ell_\infty$ | Specific | decision | 2019 |
| Wieland(Brendel et al., 2019) | Untargeted, Targeted | $\ell_0$, $\ell_2$, $\ell_\infty$ | Specific | decision | 2019 |
| BayesOpt(Ru et al., 2020) | Untargeted, Targeted | $\ell_2$ | Specific | score | 2020 |
| DFOMeunier et al. (2019) | Untargeted, Targeted | $\ell_\infty$ | Specific | score | 2020 |
| Meta Attack(Du et al., 2019) | Untargeted, Targeted | $\ell_2$ | Specific | score | 2020 |
| Attack on Attention(Chen et al., 2022a) | Untargeted, Targeted | $\ell_2$, $\ell_\infty$ | Specific | attention | 2020 |
| Aha(Li et al., 2021a) | Untargeted, Targeted | $\ell_2$ | Specific | decision | 2021 |
| Simulator Attack(Ma et al., 2021) | Untargeted, Targeted | $\ell_2$, $\ell_\infty$ | Specific | score | 2021 |
| Derivative-free(Yang and Long, 2021) | Untargeted, Targeted | $\ell_2$, $\ell_\infty$ | Specific | score | 2021 |
| data-free UAP(Zhang et al., 2021) | Untargeted, Targeted | $\ell_\infty$ | Universal | score | 2021 |

**Table 4**
Geometric-transformation-based attacks.

| Attack | Target/Untargeted | Norm | Black/White | Category | Year |
|---|---|---|---|---|---|
| ManiFool(Kanbak et al., 2018) | Untargeted | $\ell_\infty$ | White | geometric-transformation | 2018 |
| stAdv(Xiao et al., 2018) | Untargeted, targeted | $\ell_2$ | White | geometric-transformation | 2018 |
| DIM(Xie et al., 2019) | Untargeted | $\ell_\infty$ | White | gradient, geometric-transformation | 2019 |
| TIM(Dong et al., 2019a) | Untargeted, targeted | $\ell_\infty$ | White | gradient, geometric-transformation | 2019 |
| SIM(Lin et al., 2020) | Untargeted | $\ell_\infty$ | White | gradient, geometric-transformation | 2020 |
| CIM(Yang et al., 2021) | Untargeted | $\ell_2$, $\ell_\infty$ | White | geometric-transformation | 2021 |
| Admix(Wang et al., 2021a) | Untargeted | $\ell_\infty$ | White | geometric-transformation | 2021 |
| AITL(Yuan et al., 2021b) | Targeted | $\ell_\infty$ | Black | transfer, geometric-transformation | 2021 |
| AVA(Tian et al., 2021) | Targeted | $\ell_\infty$ | Black | transfer, geometric-transformation | 2021 |

the principle applied is geometric transformation invariance. The geometric-transformation-based attack steps usually need modified gradient updates and input transformations.

In Table 4, we summarize and conclude some of the geometric-transformation-based attacks. Recent studies like Lin et al. (2020); Yang et al. (2021) have obtained higher attack performance by employing more efficient geometric transformations. At the same time, Tian et al. (2021) introduced the approach of vignetting, a very natural and unobtrusive processing, to carry out the attack, gaining an excellent attack practicability. In addition, the perturbation norm of various geometric-transformation-based attacks is often $\infty$ due to the specificity of geometric changes on image modifications. Note that all attacks based on geometric transformations are specific, because the geometric transformations that cause errors in the classification of different models are not the same. Also, multiple interactions with the model are required to obtain a more efficient transformation, so all attacks in this category are iterative.

Inspired by data enhancement, Xie et al. proposed DIM (Xie et al., 2019), which improves the transferability of adversarial samples by creating diverse input patterns, ranging over flipping, rotating, cropping and scaling of images. In the actual attack process, DIM can be improved based on momentum, with MI-FGSM (Dong et al., 2018), etc. obtained good attack effect. Dong et al. (2019a) proposed a translation-invariant attack method TIM, which makes the attacked model less sensitive to the classification of the corresponding adversarial samples by translational transformation of the adversarial samples, resulting in the improved transferability of the adversarial samples. This attack is applicable to any gradient optimization attack method by performing a convolution operation before the gradient is applied to the original image. The authors also suggest that TI-DIM has the best attack performance and that the gaussian kernel is the best choice for the convolution operation.

*SIM:* Since DNNs are scaling invariant, Lin et al. (2020) improves the transferability of the adversarial samples by optimizing the adversarial perturbation on the transformed image copy, i.e., using the average gradient of the transformed image instead of the currently computed gradient. Also the authors introduce the nesterov accelerated gradient into the iterative gradient-based attack, thus effectively look forward, and improve the transferability of adversarial examples. The authors propose SI-NI-TI-DIM (Scaling Invariant, Nesterov Iterative FGSM Integrated Translational Invariance Diversity Input Method), which can achieve an average success rate of 93.1% in the black-box setting.

*Admix:* Unlike the previous input transformation based methods, Yang et al. (2021) achieves a better attack generalization capability by mixing up multiple image samples in a master-slave manner and proposing no blending of labels. Among them, Admix makes the sample points closer to the decision boundary by using information from other classes in order to obtain better gradient information to achieve counterattack. The authors perform the attack by combining admix and MI-FGSM and achieve a 5–10% improvement in the success rate of the attack on the current benchmark.

## 5. Construction of knowledge graph

Knowledge graph is a large-scale semantic network, an abstract description of the real world, by structuring heterogeneous knowledge in a domain in order to build connections between knowledge. By constructing a knowledge graph, we can observe the development in this field.

In this section, we use VOSviewer, an information visualization technique and tool for dynamic network analysis, to construct knowledge graph of the adversarial attack domain for the visualization and analysis work of the field in the later sections. First, we introduce the theoretical basis of knowledge graph construction. Then, we collect and process the required source data, and show the node structure of the knowledge graph. Moreover, we briefly define the relevant parameters in the construction of the knowledge graph.

### 5.1. Theoretical basis

We use VOSviewer based on citation analysis and co-citation analysis to create theoretical models that map from the "knowledge base" to the "research frontier" and use time-sliced snapshots to show the evolution of the research field. It can be used to detect and visualize emerging trends and sudden changes in $\Phi(t)$ over time. It is broadly defined as follows.

$$\begin{aligned}
&\Phi(t) : \Psi(t) \rightarrow \Omega(t) \\
&\Psi(t) = \{\ \text{term}\ \mid\ \text{term}\ \in S_{\text{title}} \cup S_{\text{abstract}} \cup S_{\text{descriptor}} \\
&\qquad \cup S_{\text{identifier}} \wedge\ \text{IsHotTopic}\ (\ \text{term}, t)\} \\
&\Omega(t) = \{\ \text{article}\ \mid\ \text{term}\ \in \Psi(t) \wedge\ \text{term}\ \in\ \text{article}_{\,0} \\
&\qquad \wedge\ \ \text{article}_{\,0} \rightarrow\ \text{article}\ \}
\end{aligned} \tag{5}$$

Where $\Psi(t)$ is a set of research-front terms associated with trends and emergence at time $t$, and $\Omega(t)$ consists of the set of articles cited by articles that found research-front terms. $S_{\text{title}}$ represents a set of title terms, IsHotTopic ( term, $t$ ) denotes a Boolean function, and $article_{\,0} \rightarrow article$ indicates that $article_{\,0}$ cites the *article*.

By measuring the literature (set) in the field of adversarial attacks, it is possible to explore the critical paths and knowledge turning points in this field, and to form a series of visual maps to analyze the potential dynamic mechanisms of disciplinary evolution and detect the frontiers of disciplinary development.

### 5.2. Data sources and pre-processing

According to the data source requirements of VOSviewer, Scopus was selected as the literature search engine. In the search, we selected "Adversarial attack" as the search topic, filtered the literature type as "Article" and "Survey ". The time range was selected as all. We refined and eliminated articles that were not related to the topic, and finally obtained 5923 records. The data collection time was March 7, 2022.

To meet the structure requirements of VOSviewer for the knowledge graph nodes, we check and reconstruct the data nodes. The format of the data nodes used for data information processing in VOSviewer is shown in Table 5.

### 5.3. Parameter definition

*Node size.* The size of a node reflects the size of the number of nodes referenced by other nodes. In order to ensure the reasonableness of the image information display, the node size corresponds to the number of nodes differently for different images.

*Node betweenness centrality.* The betweenness centrality is used in VOSviewer to measure the importance of a node in the network, and more connections mean that the node has a higher node centrality. In this knowledge graph building, we use the entropy cosine distance to calculate the connection strength between nodes:

$$\text{Cosine}\left(C_y, S_i, S_j\right) = \frac{C_{ij}}{\sqrt{S_i S_j}} \tag{6}$$

**Table 5**
Description of the main parameters of the data node format.

| Full Name | Description |
|---|---|
| Publication Type | Conference Article |
| Author | Carlini, N Wagner, D |
| Title | Towards Evaluating the Robustness of Neural Networks |
| Source | 2017 IEEE SYMPOSIUM ON SECURITY AND PRIVACY (SP) |
| Abstract | Neural networks provide state-of-the-art results for most machine learning tasks. Unfortunately, neural networks are vulnerable to adversarial examples: given an input x and any target classification t, it is possible to find a new input x' that is similar to x but classified as t... |
| Cooperate Location | Univ Calif Berkeley, Berkeley, CA |
| Cited Reference | Andor Daniel, 2016, ARXIV160306042 BASTANI O., 2016, ARXIV160507262 Bojarski M., 2016, END END LEARNING SEL BOURZAC K., 2016, BRINGING BIG NEURAL Carlini Nicholas, 2016, 25 USENIX SEC S USEN ... |
| Publication Year | 2017 |

**Table 6**
Co-citation index top10.

| No. | Year | Paper Title | Co-citation Index | WOS Citation Counts |
|---|---|---|---|---|
| 1 | 2017 | Towards Evaluating the Robustness of Neural Networks(Carlini and Wagner, 2017) | 371 | 1425 |
| 2 | 2016 | DeepFool: a simple and accurate method to fool deep neural networks(Moosavi-Dezfooli et al., 2016) | 255 | 1038 |
| 3 | 2016 | The Limitations of Deep Learning in Adversarial Settings(papernot et al., 2016) | 249 | 1018 |
| 4 | 2016 | Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks(Papernot et al., 2016) | 187 | 704 |
| 5 | 2017 | Practical Black-Box Attacks against Machine Learning(Papernot et al., 2017) | 180 | 768 |
| 6 | 2017 | Universal adversarial perturbations(Moosavi-Dezfooli et al., 2017) | 141 | 415 |
| 7 | 2018 | Boosting Adversarial Attacks with Momentum(Dong et al., 2018) | 132 | 317 |
| 8 | 2018 | Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey(Akhtar and Mian, 2018) | 130 | 405 |
| 9 | 2016 | Deep Residual Learning for Image Recognition(He et al., 2016) | 121 | 20,397 |
| 10 | 2019 | One Pixel Attack for Fooling Deep Neural Networks(Su et al., 2019) | 99 | 326 |

**Table 7**
Burst keyword list.

| Keywords | Strength | Begin | End | 2012–2022 |
|---|---|---|---|---|
| adversarial risk analysis | 4.24 | 2012 | 2018 | |
| scheme | 2.58 | 2012 | 2019 | |
| wireless sensor network | 2.57 | 2013 | 2019 | |
| evasion attack | 4.35 | 2014 | 2018 | |
| privacy | 3.23 | 2014 | 2019 | |
| cyber-physical system | 3.48 | 2017 | 2019 | |
| security | 9.68 | 2018 | 2020 | |
| robustness | 5.12 | 2018 | 2020 | |
| algorithm | 4.14 | 2018 | 2020 | |
| strategy | 3.77 | 2018 | 2019 | |
| malware detection | 5.73 | 2019 | 2020 | |
| computational modeling | 6.48 | 2020 | 2022 | |
| generative adversarial network | 6.14 | 2020 | 2022 | |

where $S_i$ is the frequency of occurrence of $i$, $S_j$ is the frequency of occurrence of $j$, and $C_{ij}$ is the number of co-occurrence of $i$ and $j$. Each value is normalized between 0 and 1.

*Co-citation index*. Two (or more) papers are considered to constitute a co-citation relationship when they are simultaneously cited by one or more subsequent papers, and frequent co-citations indicate that they share a relevant research topic (Small, 1973). In the calculation of the co-citation index in this paper, if literature A cites both C and D, then C and D are co-cited, and the number of documents citing both is called co-citation intensity, which is 1. If literature A and B cite C, D and E, then C, D and E are co-cited, and the co-citation intensity is 2, and so on. The co-citation index of each node is the sum of all co-citation intensities of that node, and we used it in Table 6. The co-citation relationship of literature changes over time, and the development and evolutionary dynamics of a discipline can be explored through literature co-citation network studies.

*WOS citation counts*. We also refer to the number of citations in the literature search results in Web of Science for additional analysis of the importance of the literature to improve the reference value of the relevant literature, and we use it in Table 6.

*Keyword bursty strength*. We introduce keyword emergence detection to detect large changes in the number of citations at a certain time, to find the decline or rise of a particular term or keyword. For the calculation of keyword emergence strength, we uses the Klein-berg algorithm (Kleinberg, 2003), which we used in Table 7.

Based on the above theoretical basis of disciplinary development paths and standard nodal manipulation of literature data, we visualize and analyze the development in the field of counterattack by constructing a knowledge graph in Section 6. Through the establishment of the above knowledge graphs, we can more conveniently understand the development in the field of adversarial attacks.

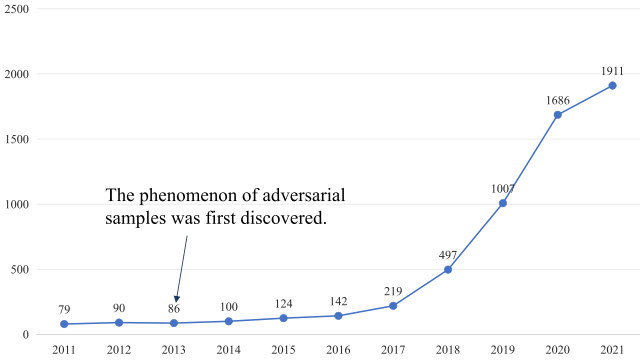



**Fig. 4.** The number of articles published in the field of adversarial attacks in 2011–2021.

## 6. Field visualization

In this section, we perform graph construction for the adversarial attack field based on the law of knowledge graph construction in the above section, and visualize and trend analysis for the corresponding graphs. Specifically, we 1) analyze the publication of papers covering time, geography, and publication distribution 6.1, 2) show institution and author collaborations 6.2, and 3) summarize the more instructive and influential papers in the field 6.3. In the figures, different node colors represent different clusters in which objects in the same cluster are similar to each other.

### 6.1. Article publication analysis

In this section, we visualize and analyze the publication of articles related to the field of adversarial attacks based on the distribution of publications in time 6.1.1, the distribution of publication regions in source 6.1.2, and the distribution of publication journals in source 6.1.3. By getting an overview of the publication situation in the field, we can effectively and intuitively grasp the overall development in the field of adversarial attacks.

#### 6.1.1. Time distribution

Since the number of publications in the field of adversarial attacks was less than ten before 2010, for the sake of clear data display, only articles after 2011 were analyzed. The distribution of research papers in the field of adversarial attacks in time is shown in Fig. 4. The phenomenon of adversarial sample was first discovered and proposed (Szegedy et al., 2014) in 2013, which triggered a strong research enthusiasm among researchers. From 2013 to 2017, works about adversarial attacks had a slow but steady publication and field development. After 2017, the field research entered a new phase of development, manifested in a strong rise in research enthusiasm. The research work in this field has shown a rapid growth trend and has not yet reached its peak. Based on the number of ar-

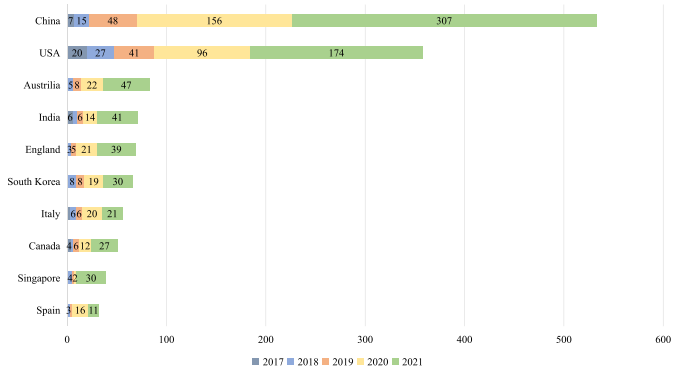



**Fig. 5.** The number of articles published by countries.

ticles published since 2022, the number of published studies in the field of adversarial attacks will reach more than two thousand in 2022. It is evident that adversarial attack research is still a great research hotspot at present.

#### 6.1.2. Geographical distribution

Figure 6 presents the distribution of the field of adversarial attacks and the cooperation relationship between scholars from different countries, which was made using the visualization tool Biblioshiny (Aria and Cuccurullo, 2017) because of its preeminence in geographic visualization. It involves 87 countries and regions, and the connecting lines represent the cooperation relationship between countries. The color represents the number of documents by the country, and the most number of articles issued is China, with 578 articles. The connections between nodes represents the centrality of country studies, and it can be seen that the United States has obtained a higher centrality with fewer publications (481), i.e., the United States has the most collaborations with other countries. In addition, countries such as Australia (90), India (84), Italy (81), South Korea (77), England (77), Canada (62), Spain (51), and Singapore (50) have more than 50 publications as shown in Fig. 5, which shows the article distribution and proportion of articles in different countries.

Based on the geographical distribution, it can be seen that China and the United States have the leading position in terms of the number of publications in this field, and there is a considerable amount of research in the field of adversarial attacks in all major countries and regions in the world, which proves that adversarial attacks and the corresponding security research work are widely valued worldwide; research in this field have not become difficult for national boundaries, and the cooperation between countries is frequent and intensive. The above shows that the relevant research has a good internationalization, which makes the academic research in the field develop in a healthy way.

#### 6.1.3. Published sources distribution

The source of articles indicates the publication situation of the literature published about adversarial attacks in different publications. Due to the rapidly evolving character of the field of adversarial attacks, unlike the publication of articles in other fields, many articles related to the field are published in conferences as a way to present the research results of scholars in a timely manner and to promote academic communication in the field. Therefore, in this analysis of publication sources, we will combine journal and conference together. After combining the conferences, we use VOSviewer to construct a network of publication sources. We filtered journals with more than 5 publications in total and more than 300 total citations, and finally obtained 23 publication results, as shown in Fig. 7, where the size of the node indicates the influence strength of that node.

We can see that conference-type publication sources have an important place and role in the field of adversarial attacks, with high-impact conferences including CVPR (h5[2]:356), ICLR (h5:253), NIPS (h5:245), ICML (h5:204), ICCV (h5:197), AAAI (h5:157), IJCAI (h5:105), KDD (h5:104) and so on. It can be seen that conferences that are very important in the field of computing are also important for the field of adversarial attacks. In addition, classical journals such as Information Sciences (h5:113), Pattern Recognition (h5:99), IEEE Transactions On Information Forensics And Security (h5:92), IEEE Transactions On Dependable And Secure Computing (h5:59) have also become important publications in this field. In total, there are 1651 articles published in journals and 4143 papers published in conferences (containing 240 reviews). Meanwhile, the conference has newer publications in the field of adversarial attacks, as seen in the colors.

### 6.2. Collaboration network analysis

We visualize and analyze the collaboration networks concerning adversarial attacks in this section, containing the institution collaboration network 6.2.1 and the author collaboration network 6.2.2. The visual analysis of institution and author collaboration networks in the field allows us to grasp the collaboration preferences and difference analysis among different institutions and authors, and use it to understand the research segmentation of related institutions and teams.

#### 6.2.1. Institution collaboration network

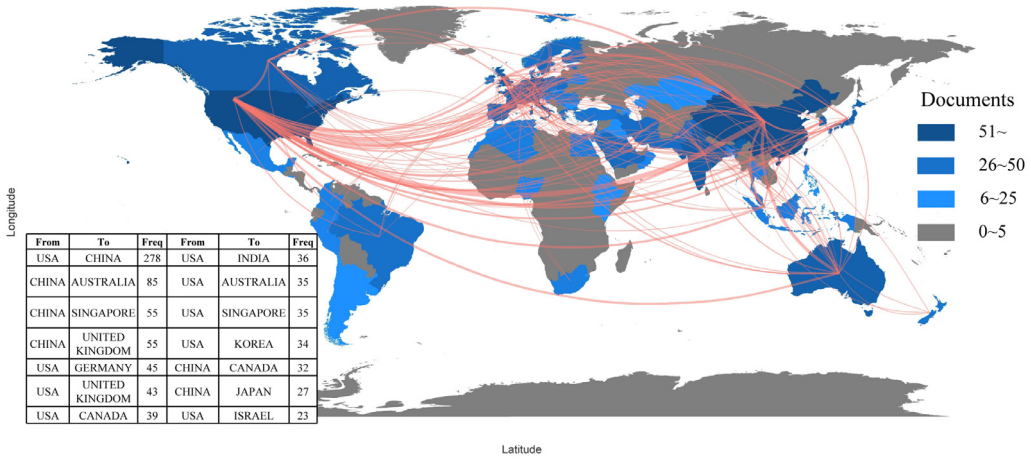
After filtering, a total of 1105 research institutions appear in the field of adversarial attacks. To obtain a clearer picture of the institution collaboration network, we refine the nodes that appear in the institution collaboration network, making the thresholds of the minimum number of articles is 5 as well as the minimum number of citations is 30. Fig. 8 presents the distribution of the network of research institutions in the field and their collaboration information, which covers 82 research institutions.

The three research institutions with the highest number of publications in the field are Chinese Acad. Sci., Zhejiang Univ., and Guangzhou Univ., with 35, 31, and 26 publications, respectively. It can be seen that there are a considerable number of research institutions in the field of adversarial attacks, and there is no academic monopoly; the cooperation links between different institutions are extensive and intensive, and the enthusiasm of cooperation between institutions is high.

#### 6.2.2. Author collaboration network

After filtering, a total of 10,124 authors appear in the field of adversarial attacks. To obtain a clearer picture of the author collaboration network, we refine the nodes that appear in the author collaboration network, making the thresholds of the minimum number of articles is 2 and the minimum number of citations is 350. Finally, we obtain the author collaboration network as shown in Fig. 9, which embraces 112 authors with 11 clusters. The scholars with more publications include Chen, Biggio, Tondi, Roli, Guizani, Du, Hyun, Choi, Zheng, Hyunson etc., all published more than 8 papers, but only accounted for 4.83% of all authors with 2 or more papers, and the difference in the number of papers published between each other was not significant. In terms of the number of authors cited, Wagner, Carlini, Biggio, Roli, Li etc. have a higher influence in the field, with more than 2000 citations.

[2] H5-index is the h-index for articles published in the last 5 complete years. Source: google scholar.




| From | To | Freq | From | To | Freq |
|---|---|---|---|---|---|
| USA | CHINA | 278 | USA | INDIA | 36 |
| CHINA | AUSTRALIA | 85 | USA | AUSTRALIA | 35 |
| CHINA | SINGAPORE | 55 | USA | SINGAPORE | 35 |
| CHINA | UNITED KINGDOM | 55 | USA | KOREA | 34 |
| USA | GERMANY | 45 | CHINA | CANADA | 32 |
| USA | UNITED KINGDOM | 43 | CHINA | JAPAN | 27 |
| USA | CANADA | 39 | USA | ISRAEL | 23 |

**Fig. 6.** Geographical distribution of field cooperation relations.

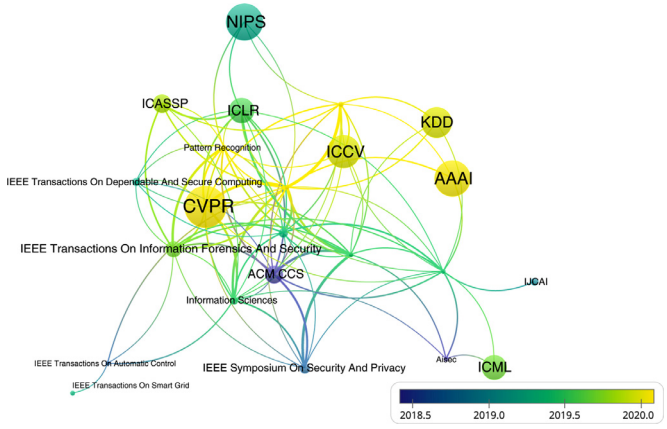



**Fig. 7.** Publication Source Network.

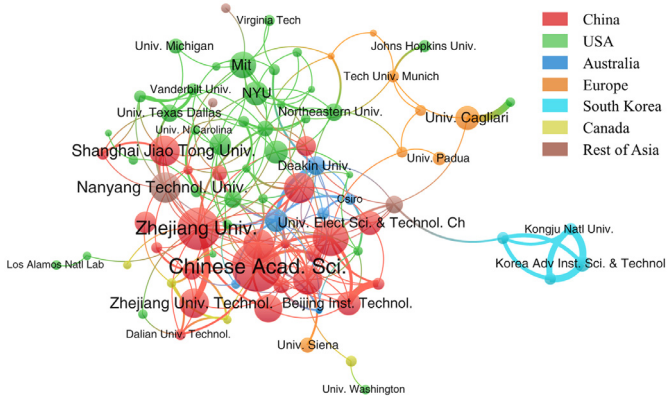



**Fig. 8.** Distribution of research institutions.

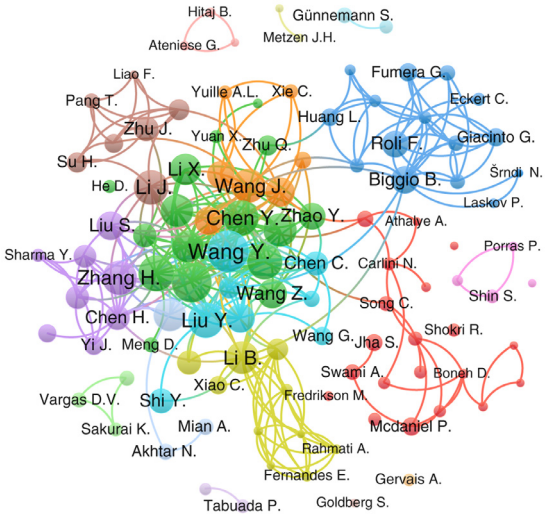



**Fig. 9.** Author Collaboration Network.

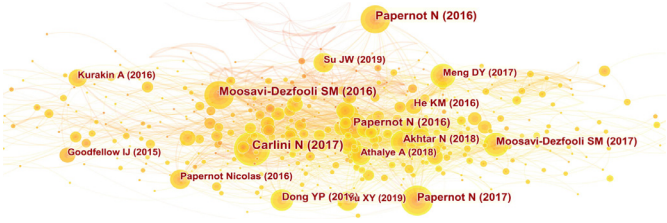



**Fig. 10.** Co-cited article network.

### 6.3. Key article analysis

We provide a visual analysis of key articles pertaining to adversarial attacks in this section. By summarizing and analyzing key chapters in the development of the field, we can better understand the key research content and academically important nodes in the field as a way to provide relevant researchers with an valuable supplement to the background of the field.

Figure 10 presents the co-citation of key articles in the field. The co-citation refers to two (or more) papers being cited by one or more later papers at the same time. The size of the node represents the frequency of the cited literature, and the larger the node, the more frequent it is. The literature with more occurrences represents the research focus of the field. In order to keep the relationship between the nodes in the network clear, only the nodes with the highest frequency are marked.

The analysis shows that there is a high number of literature with a high number of citations. Papernot et al. (2016) proposed target attack JSMA, which uses Jacobi matrices to compute a significant graph from input to output to achieve misclassification by modifying only a small number of input values. They (Papernot et al., 2016) also make the model more robust to perturbations by using the knowledge extracted during distillation to reduce the magnitude of the adversarial samples generated by the

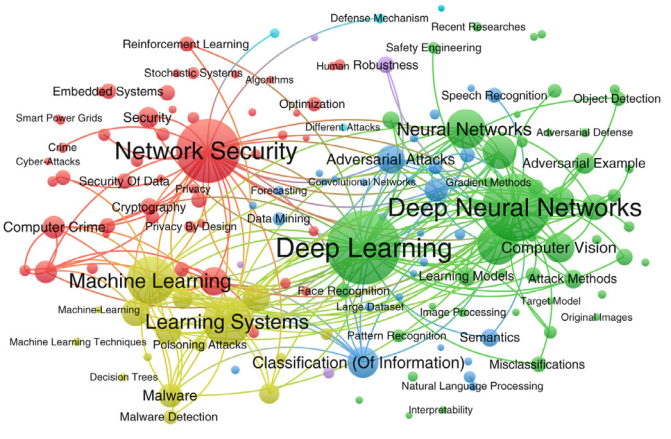



**Fig. 11.** Keyword Network.

attacker using the gradient. Carlini and Wagner (2017) demonstrate that defensive distillation does not significantly improve the robustness of the model and propose a superior effect targeted attack approach C&W. Moosavi-Dezfooli et al. (2016) proposed the untargeted attack algorithm DeepFool, whose goal is to seek minimal perturbations to achieve the goal of generating adversarial samples. Papernot et al. (2017) implemented black-box attacks by generating substitute models to simulate the decision boundaries of the approximated attacked model, etc.

Table 6 shows the top ten articles with co-citation index, the calculation of which is described in detail in Section 5.3.

## 7. Trends and directions

In this section, we present some research trends and research directions in the area of adversarial attacks based on the results of keyword dwtection analysis.

### *7.1. Trends by keyword detection analysis*

We visualize and analyze keywords in the field of adversarial attacks in this section, with the field keyword network and emergence, and keywords based on the attack-related papers under each taxonomy. By analyzing the keywords in the field, we can visualize the key content of development under this field. The keyword analysis of attack literature based on taxonomy allows us to grasp the key trends and directions researchers are focusing on in their attack exploration.

#### *7.1.1. Field keywords*

We performed keyword detection analysis on the development regarding adversarial attack, and the generated keyword citation distribution is shown in Fig. 11, containing 6 clusters and 153 keywords. From the keyword citation distribution, we can understand the hot content of research in the field.

The green cluster represented by the keywords “Deep Learning”, “Computer Vision” and “Attack Model” show the theoretical basis and mathematical models for adversarial attacks, which indicate that the development of the whole field is based on the framework of deep neural network learning, and focuses more on computer vision. It can be found that image processing-based tasks (“Object Detection” and “Face Recognition”) are important hotspots in the field. The red cluster represented by the keywords “Network Security”, “Computer Crime”, and “Privacy By Design” show the research objectives in the field of adversarial attacks. The red cluster shows the research goals about adversarial attacks, which are to enhance the application value of deep neural networks in security-critical areas and to reduce the damage caused by malicious perturbations. In addition, the red cluster also indicates the importance of system design in the field. The yellow cluster represented by the keywords “Machine Learning”, “Learning Systems”, and “Training Data” show the importance of learning with regards to adversarial attacks, which is reflected in the fact that the conduct of adversarial attacks improves the robustness of the model, by ways of learning, and depending on the quality of the training data. The blue cluster represented by “Classification”, “Convolutional Neural Networks”, and “Forecasting” reflect that in the current adversarial learning field, the models used are mainly structured as convolutional neural networks, and the tasks of the models are mainly focused on classification and prediction tasks, which is in line with the development characteristics of this field. The blue cluster also includes the keywords “performance” and “Benchmark Datasets”, showing the importance of being able to compare the attack performance of the models in the evaluation of the experimental training effects. Purple clustering, represented by the keywords “Robustness”, “Classification Accuracy”, and “Big Data”, shows the task characteristics in the field of adversarial attacks, including better measuring the robustness of a model, improving the classification accuracy of a model, and improving the ability of a model to cope with potential security problems in big data. The cyan clustering represented by the keywords “Different Attacks” and “Defense Strategy” shows the breadth of the development of adversarial attacks. By continuously exploring more types of attacks to better discover the defects of the model, and by proposing corresponding defense strategy in the face of different attacks, the robustness of the model is improved.

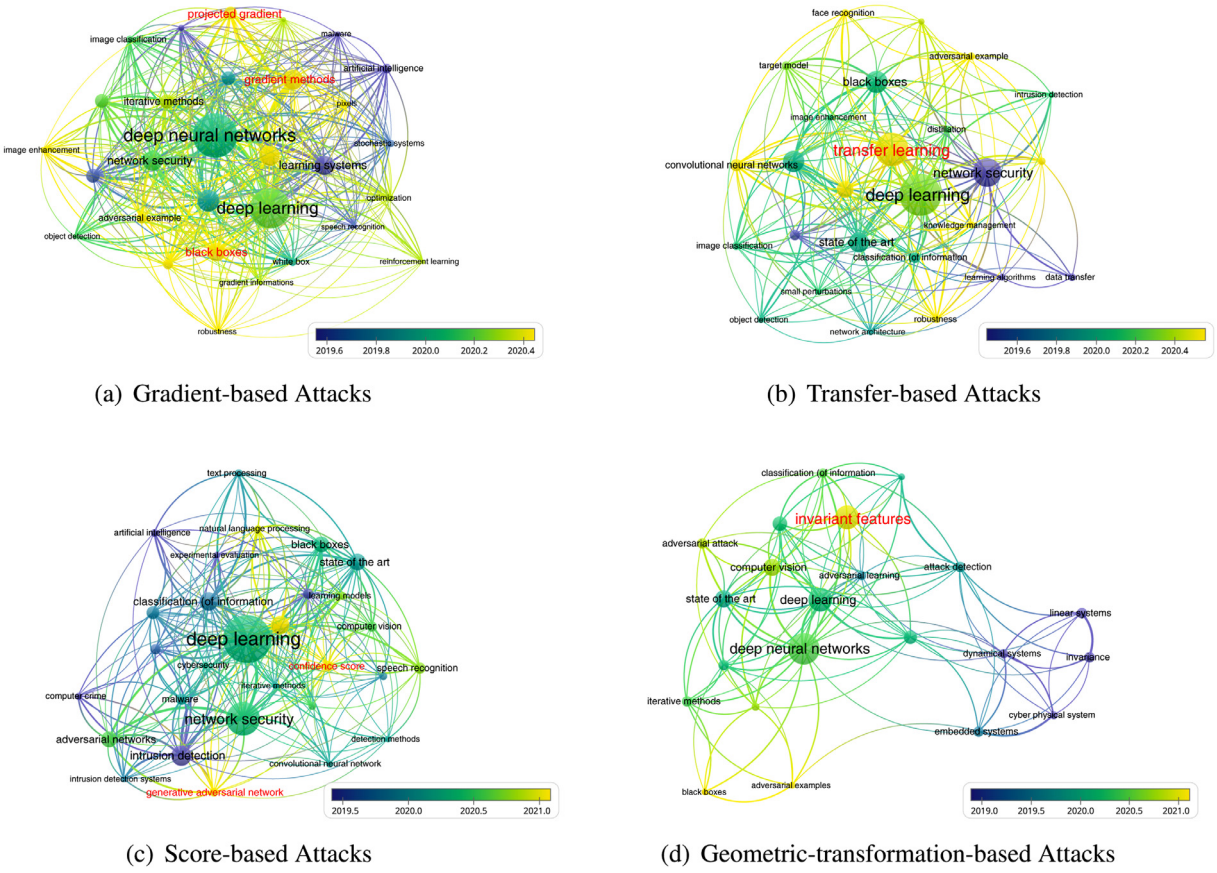
In addition, we apply keyword emergence detection methods to evaluate the content of research hotspots with emergence in the field of adversarial attacks. Table 7 presents the analysis of keyword burst in the adversarial attack domain. We uses the Kleinberg (Kleinberg, 2003) algorithm to detect word frequency. This algorithm was proposed by Kleinberg, which is based on text data mining technology and aims to discover the sudden increase of a certain research direction in a certain research field over a period of time.

As we can see from Table 7, the evaluation of the risk of the model, the enhancement of security and the protection of privacy are important goals of the development in the field of counterattack in recent years. Meanwhile, research topics targeting cyber-physical system and malware detection are more novel beyond the field of image. In addition, the introduction of generative adversarial network and the construction of computational model has become a hot issue in research in the past two years in order to obtain better attack effects.

#### *7.1.2. Taxonomy based attack strategy keywords*

Taxonomy-based attack strategy research is a more common and systematic approach in the field of adversarial attacks. In this section, we use the visual analysis of keyword graphs of taxonomy-based attack strategies to explore the hotspots of field research under different attack taxonomies. Among them, the taxonomy of attacks mainly includes 1) gradient-based attack, 2) transfer-based attack, 3) score-based attack and 4) geometric- transformation-based attack. In the following, we will explore the hotspots of attack strategies based on the above taxonomy through Figure12, where the nodes in yellow have the latest average year of publication.

Figure 12(a) shows the keyword distribution of the gradient-based attack study. From the figure we can see that the attack strategy focuses more on the gradient strategy, which achieves better attack performance through a better gradient optimization strategy. Meanwhile, in the process of using gradient, the more effective way is to use projected gradient. It is experimentally proven




(a) Gradient-based Attacks

(b) Transfer-based Attacks

(c) Score-based Attacks

(d) Geometric-transformation-based Attacks

**Fig. 12.** Keyword networks of adversarial attacks based on taxonomy.

Madry et al. (2018) that this way can obtain excellent model performance. In addition, gradient-based attacks are usually considered as white-box attacks, but in recent research scholars are more willing to explore attacks in black-box settings. Better black-box attack effects have better practical applications. Fig. 12(b) shows the distribution of keywords for the transfer-based attack study. From the figure, we can observe that transfer learning is an important step in this kind of attack strategy, namely, by way of building a learning system. Moreover, in addition to transferring for models, scholars also focus on transferring strategies about data, such as original data before adding perturbation and adversarial samples after perturbation. Moreover, transfer-based attack focuses more on knowledge management and shows strong interest in large datasets. Fig. 12(c) shows the distribution of keywords for the score-based attack study. From the figure we can see that the confidence score is the keyword that appears more often in comparison, which is due to the implementation principle of this attack strategy. In order to interfere with the confidence score of higher mislabeling, the iterative method is also focused on. In addition, to reduce the number of queries to the target model, score-based attacks often employ the structure of generative adversarial network to train out substitute models. Fig. 12(d) shows the distribution of keywords for the geometric-transformation-based attack study. From the figure, we can see that finding invariant features becomes the focus of this attack strategy. A more effective and practical geometric transformation is used to achieve a more enhanced attack performance. In addition, inspired by enhancement learning, scholars also adopt the strategy of image enhancement to enlarge the training set of the attack model, which can correspondingly improve the robustness of the target model.

### 7.2. Research directions

Although there has been great progress in the field of adversarial attacks, we can still find many research trends and aspects worthy of attention in the above analysis. On the one hand, in terms of models, better attack accuracy, transfer capability, and black-box capability become the main directions for model improvement. On the other hand, in order to enable further applications of deep neural networks in security critical areas, the models need to have better robustness, by implementing more effective defense strategies, etc. In addition, more application scenarios, neural network types should also be focused.

*New application scenarios.* As shown in the above discussion, adversarial attacks have gained more progress in scenario of image classification and recognition. Moreover, in the analysis of keyword detection in the field, it can be found that a considerable number of scholars are already focusing on application scenarios beyond images. Therefore, we can start from the study of images and gradually find the application points of adversarial attacks in Graph, NLP and video and other related scenarios, based on the mature techniques of adversarial attacks in the image domain. By continuously exploring new application scenarios, we can obtain the robustness of the model in different scenarios as a way to improve the application value of deep neural networks in security critical areas. *New networks for attacks.* Convolutional neural network shows powerful performance in classification and prediction tasks in the image domain, and thus CNN-based adversarial attacks are currently the mainstream of field development. In addition to the network structure of CNN, there are also other types of network structures, including recurrent neural network (RNN), neural networks for reinforcement learning, etc. However, there is

less research work on adversarial attacks for other types of network structures, and the diversity of research in the field of adversarial attacks in terms of network types is lacking. Therefore, starting from networks used for image classification, related researchers can focus more on network types such as RNN and neural networks in reinforcement learning to improve the security and robustness of different neural network types. *More efficient adversarial attack networks.* In Szegedy et al. (2014), the authors first turned the process of finding adversarial perturbations into solving convex optimization problems. Since then, many researchers have used different techniques to apply in convex optimization as a way to obtain better attack efficiency in adversarial attack networks. Starting from primitive optimization methods, to using strategies such as meta learning to optimize queries for adversarial networks. Despite the tremendous progress in current research work, there are still many attack strategies that fail to achieve the desired attack effect for different network types, different model settings, and different quantities of resources provided. Therefore, researchers can look for newer techniques to be applied to convex optimization to enhance the attack efficiency of adversarial attack networks in different settings. *Adversarial defense strategies.* The proposal of various efficient attack strategies represents more security vulnerabilities about deep neural networks are discovered. To enhance the security of the model, each attack method has a corresponding counter defense strategy. Thus, the security of neural networks is improved by proposing more effective adversarial defense strategies to resist more enhanced attack methods. *Generative Adversarial Network (GAN).* The introduction of GAN has brought a great deal of research interest to the community and has produced different variants to suit different application settings. With GAN networks, inputting known data, computers can learn and create completely new synthetic data and use it to train more effective models. Based on the innovative features of GAN, by combining GAN with adversarial attack research, it is possible to simultaneously construct more effective adversarial samples and implement defense strategies that make the model more robust. *Evaluation.* One of the goals of adversarial samples is to improve the robustness of DNNs. The robustness of a model can be assessed by measuring the effectiveness of the attack on the model. Therefore, looking for an evaluation method that accurately assesses the effectiveness of an attack method or defense strategy can better measure the robustness of a model.

## 8. Conclusion

The adversarial sample phenomenon has become a non-negligible obstacle for the application of deep learning networks in safety-critical areas. In this paper, we provide an in-depth and comprehensive review of adversarial attacks in the field of computer vision. We summarize the attack strategies based on a refined taxonomy. Also, to better explore the development status of the adversarial attacks field, we visualize and analyze the field literature using knowledge graphs and conclude field trends using keyword detection. Research directions with value and practical implications are proposed based on a comprehensive field analysis from various perspectives such as model improvement and application.

In this paper, we intend to provide researchers in computer vision with guidance on the research of adversarial attacks. However, due to the limitations of the research scope, newly proposed adversarial attacks in scenarios such as natural language processing, graphs, speech recognition, etc. have not been reviewed and visually analyzed, which awaits more future work to complete.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The authors would like to thank the reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China under Grants No. 62002332, 62072443.

## References

Akhtar, N., Mian, A., Kardan, N., Shah, M., 2021. Advances in adversarial attacks and defenses in computer vision: a survey. IEEE Access 9, 155161–155196.
Akhtar, N., Mian, A., Kardan, N., Shah, M., 2021. Threat of adversarial attacks on deep learning in computer vision: survey II. arXiv e-prints arXiv–2108.
Akhtar, N., Mian, S.A., 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 14410.0–14430.0.
Alaifari, R., Alberti, S.G., Gauksson, T., 2019. ADEF: an iterative algorithm to construct adversarial deformations. international conference on learning representations.
Aria, M., Cuccurullo, C., 2017. Bibliometrix: an r-tool for comprehensive science mapping analysis. J. Informetrics 11 (4), 959–975.
Aryal, K., Gupta, M., Abdelsalam, M., 2021. A survey on adversarial attacks for malware analysis. arXiv preprint arXiv:2111.08223.
Brendel, W., Rauber, J., Bethge, M., 2018. Decision-based adversarial attacks: reliable attacks against black-box machine learning models. international conference on learning representations.
Brendel, W., Rauber, J., Kümmerer, M., Ustyuzhaninov, I., Bethge, M., 2019. Accurate, reliable and fast robustness evaluation. Advances in Neural Information Processing Systems 32 (NIPS 2019) 12841–12851.
Carlini, N., Wagner, D.A., 2017. Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy. IEEE Computer Society, pp. 39–57.
Chen, C., 2006. Citespace ii: detecting and visualizing emerging trends and transient patterns in scientific literature. JASIST 359–377.
Chen, C., Hu, Z., Liu, S., Tseng, H., 2012. Emerging trends in regenerative medicine: a scientometric analysis in citespace. Expert Opin Biol Ther 593–608.
Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.-J., 2018. EAD: elastic-net attacks to deep neural networks via adversarial examples. national conference on artificial intelligence.
Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.-J., 2017. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. AISec@CCS 15–26.
Chen, S., He, Z., Sun, C., Yang, J., Huang, X., 2022. Universal adversarial attack on attention and the resulting dataset damagenet. IEEE Trans Pattern Anal Mach Intell 2188–2197.
Chen, T., Ling, J., Sun, Y., 2022. White-box content camouflage attacks against deep learning. Computers & Security 117, 102676. doi:10.1016/j.cose.2022.102676.
Cheng, S., Dong, Y., Pang, T., Su, H., Zhu, J., 2019. Improving black-box adversarial attacks with a transfer-based prior. Advances in Neural Information Processing Systems 32 (NIPS 2019) 10932–10942.
Croce, F., Hein, M., 2019. Sparse and imperceivable adversarial attacks. 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019) 4723–4731.
Ding, J., Xu, Z., 2020. Adversarial attacks on deep learning models of computer vision: a survey. international conference on algorithms and architectures for parallel processing 396–408.
Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J., 2018. Boosting adversarial attacks with momentum. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 9185–9193.
Dong, Y., Pang, T., Su, H., Zhu, J., 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019) 4307–4316.
Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J., 2019. Efficient decision-based black-box adversarial attacks on face recognition. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019) 7706–7714.
Du, J., Zhang, H., Zhou, T.J., Yang, Y., Feng, J., 2019. Query-efficient meta attack to deep neural networks. CoRR.
Eck, J.v.N., Waltman, L., 2010. Software survey: vosviewer, a computer program for bibliometric mapping. Scientometrics 523–538.
Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D., 2018. Robust physical-world attacks on deep learning visual classification. In: CVPR. Computer Vision Foundation / IEEE Computer Society, pp. 1625–1634.
Finlay, C., Pooladian, A.-A., Oberman, M.A., 2019. The logbarrier adversarial attack: making effective use of decision boundary information. 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019) 4861–4869.
Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In: CVPR. IEEE Computer Society, pp. 3354–3361.

Goodfellow, J.I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. international conference on learning representations.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: CVPR. IEEE Computer Society, pp. 770–778.
Helmstaedter, M., Briggman, K.L., Turaga, S.C., Jain, V., Seung, H.S., Denk, W., 2013. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500 (7461), 168–174.
Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29 (6), 82–97.
Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861.
Huang, Q., Katsman, I., Gu, Z., He, H., Belongie, J.S., Lim, S.-N., 2019. Enhancing adversarial example transferability with an intermediate level attack. 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019) 4732–4741.
Huang, Z., Zhang, T., 2020. Black-box adversarial attack with transferable model-based embedding. ICLR.
Kanbak, C., Moosavi-Dezfooli, S.-M., Frossard, P., 2018. Geometric robustness of deep networks: analysis and improvement. 2018 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR) 4441–4449.
Kleinberg, M.J., 2003. Bursty and hierarchical structure in streams. Data Min. Knowl. Discov. 91–101.
Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z., Li, F., 2021. A survey on adversarial attack in the age of artificial intelligence. Wireless Communications and Mobile Computing 2021.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114.
Kurakin, A., Goodfellow, J.I., Bengio, S., 2017. Adversarial examples in the physical world. international conference on learning representations.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.
Li, J., Ji, R., Chen, P., Zhang, B., Hong, X., Zhang, R., Li, S., Li, J., Huang, F., Wu, Y., 2021. Aha! adaptive history-driven attack for decision-based black-box models. ICCV 2021.
Li, Q., Fu, L., Wang, X., Zhou, C., 2021. Scientific x-ray. CoRR abs/2108.03458.
Li, Y., Cheng, M., Hsieh, C.-J., Lee, T.C., 2022. A review of adversarial attack and defense for classification methods. Am Stat 1–17.
Liang, B., Li, H., Su, M., Li, X., Shi, W., Wang, X., 2021. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Trans. Dependable Secur. Comput. 18 (1), 72–85.
Lin, J., Song, C., He, K., Wang, L., Hopcroft, E.J., 2020. Nesterov accelerated gradient and scale invariance for adversarial attacks. ICLR.
Liu, Y., Chen, X., Liu, C., Song, D., 2017. Delving into transferable adversarial examples and black-box attacks. ICLR.
Liu, Y., Moosavi-Dezfooli, S.-M., Frossard, P., 2019. A geometry-inspired decision-based attack. 2019 IEEE/CVF International Conference on Computer vision (ICCV 2019) 4889–4897.
Lu, J., Issaranon, T., Forsyth, A.D., 2017. Safetynet: detecting and rejecting adversarial examples robustly. 2017 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV) 446–454.
Ma, C., Chen, L., Yong, J.-H., 2021. Simulating unknown target models for query-efficient black-box attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11835–11844.
Machado, R.G., Silva, E., Goldschmidt, R.R., 2023. Adversarial machine learning in image classification: a survey toward the defender's perspective. ACM Comput Surv 1–38.
Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks. international conference on learning representations.
McLaughlin, N., Martinez del Rincon, J., Kang, B., Yerima, S., Miller, P., Sezer, S., Safaei, Y., Trickel, E., Zhao, Z., Doupé, A., et al., 2017. Deep android malware detection. In: Proceedings of the seventh ACM on conference on data and application security and privacy, pp. 301–308.
Meunier, L., Atif, J., Teytaud, O., 2019. Yet another but more efficient black-box adversarial attack: tiling and evolution strategies. arXiv preprint arXiv:1910.02244.
Meyer, M., Grant, K., Morlacchi, P., Weckowska, D., 2014. Triple helix indicators as an emergent area of enquiry: a bibliometric perspective. Scientometrics 151–174.
Mingyi, Z., Jing, W., Yipeng, L., Shuaicheng, L., Ce, Z., 2020. Dast: data-free substitute training for adversarial attacks. 2020 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR) 231–240.
Modas, A., Moosavi-Dezfooli, S.-M., Frossard, P., 2019. Sparsefool: a few pixels make a big difference. 2019 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2019) 9079–9088.
Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P., 2017. Universal adversarial perturbations. In: CVPR. IEEE Computer Society, pp. 86–94.
Moosavi-Dezfooli, S., Fawzi, A., Frossard, P., 2016. Deepfool: A simple and accurate method to fool deep neural networks. In: CVPR. IEEE Computer Society, pp. 2574–2582.
Papernot, N., McDaniel, P.D., Goodfellow, J.I., Jha, S., Celik, B.Z., Swami, A., 2017. Practical black-box attacks against machine learning. AsiaCCS 506–519.
papernot, n., mcdaniel, p., jha, s., fredrikson, m., celik, b.z., swami, a., 2016. The limitations of deep learning in adversarial settings. 1ST IEEE EUROPEAN SYMPOSIUM ON SECURITY AND PRIVACY 372–387.
Papernot, N., McDaniel, P.D., Wu, X., Jha, S., Swami, A., 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy. IEEE Computer Society, pp. 582–597.
Persson, O., Danell, R., Schneider, J.W., 2009. How to use bibexcel for various types of bibliometric analysis. Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday 5, 9–24.
Pham, H., Dai, Z., Xie, Q., Le, Q.V., 2021. Meta pseudo labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11557–11568.
Posada, J., Zorrilla, M., Dominguez, A., Simões, B., Eisert, P., Stricker, D., Rambach, J.R., Döllner, J., Guevara, M., 2018. Graphics and media technologies for operators in industry 4.0. IEEE Comput Graph Appl 38 (5), 119–132.
Qiu, S., Liu, Q., Zhou, S., Wu, C., 2019. Review of artificial intelligence adversarial attack and defense technologies. APPLIED SCIENCES-BASEL.
Rony, J., Granger, E., Pedersoli, M., Ben Ayed, I., 2021. Augmented lagrangian adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7738–7747.
Rony, J., Hafemann, G.L., Oliveira, S.L., Ayed, B.I., Sabourin, R., Granger, E., 2019. Decoupling direction and norm for efficient gradient-based l-2 adversarial attacks and defenses. 2019 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2019) 4317–4325.
Ru, B., Cobb, A., Blaas, A., Gal, Y., 2020. Bayesopt adversarial attack. ICLR.
Sarkar, S., Bansal, A., Mahbub, U., Chellappa, R., 2017. Upset and angri: breaking high performance image classifiers. arXiv preprint arXiv:1707.01159.
Serban, A.C., Poll, E., 2018. Adversarial examples - a complete characterisation of the phenomenon. CoRR abs/1810.01185.
Shaoxiong, J., Shirui, P., Erik, C., Pekka, M., S, P.Y., 2022. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst 494–514.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR.
Small, H., 1973. Co-citation in the scientific literature: a new measure of the relationship between two documents. Journal of the American Society for information Science 24 (4), 265–269.
Su, J., Vargas, V.D., Sakurai, K., 2019. One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. 828–841.
Sun, L., Dou, Y., Yang, C., Wang, J., Yu, P.S., He, L., Li, B., 2018. Adversarial attack and defense on graph data: asurvey. arXiv preprint arXiv:1812.10528.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: NIPS, pp. 3104–3112.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: CVPR. IEEE Computer Society, pp. 1–9.
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R., 2014. Intriguing properties of neural networks. ICLR (Poster).
Tian, B., Juefei-Xu, F., Guo, Q., Xie, X., Li, X., Liu, Y., 2021. AVA: adversarial vignetting attack against visual recognition. In: IJCAI. ijcai.org, pp. 1046–1053.
Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, D.P., 2018. Ensemble adversarial training: attacks and defenses. ICLR.
Tu, C.-C., Ting, P.-S., Chen, P.-Y., Liu, S., Zhang, H., Yi, J., Hsieh, C.-J., Cheng, S.-M., 2019. Autozoom: autoencoder-based zeroth order optimization method for attacking black-box neural networks. national conference on artificial intelligence.
Wang, J., Dong, G., Sun, J., Wang, X., Zhang, P., 2019. Adversarial sample detection for deep neural network through model mutation testing. In: ICSE. IEEE / ACM, pp. 1245–1256.
Wang, W., Wang, R., Wang, L., Wang, Z., Ye, A., 2019. Towards a robust deep neural network in texts: asurvey. arXiv preprint arXiv:1902.07285.
Wang, X., He, K., 2021. Enhancing the transferability of adversarial attacks through variance tuning. CVPR 1924–1933.
Wang, X., He, X., Wang, J., He, K., 2021. Admix: Enhancing the transferability of adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16158–16167.
Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., Ren, K., 2021. Feature importance-aware transferable adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7639–7648.
Wang, Z.-M., Gu, M.-T., Hou, J.-H., 2019. Sample based fast adversarial attack method. Neural Processing Letters 2731–2744.
Wu, W., Su, Y., Chen, X., Zhao, S., King, I., Lyu, R.M., Tai, Y.-W., 2020. Boosting the transferability of adversarial samples via attention. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 1158–1167.
Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., Song, D., 2018. Spatially transformed adversarial examples. ICLR.
Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, L.A., 2019. Improving transferability of adversarial examples with input diversity. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019) 2725–2734.
Yang, B., Zhang, H., Zhang, Y., Xu, K., Wang, J., 2021. Adversarial example generation with adabelief optimizer and crop invariance. arXiv preprint arXiv:2102.03726.
Yang, R., Long, T., 2021. Derivative-free optimization adversarial attacks for graph convolutional networks. PeerJ Comput. Sci. 7, e693.
Yao, Z., Gholami, A., Xu, P., Keutzer, K., Mahoney, W.M., 2019. Trust region based adversarial attack on neural networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019) 11342–11351.
Yu, M., Sun, S., 2022. Fe-dast: fast and effective data-free substitute training for black-box adversarial attacks. Computers & Security 113, 102555. doi:10.1016/j.cose.2021.102555.

Yuan, Z., Zhang, J., Jia, Y., Tan, C., Xue, T., Shan, S., 2021. Meta gradient adversarial attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7748–7757.
Yuan, Z., Zhang, J., Shan, S., 2021. Adaptive image transformations for transfer-based adversarial attack. arXiv preprint arXiv:2111.13844.
Zeiler, D.M., Fergus, R., 2014. Visualizing and understanding convolutional networks. COMPUTER VISION - ECCV 2014, PT I 818–833.
Zha, D.-S., Feng, T.-T., Gong, X.-L., Liu, S.-Y., 2021. When energy meets blockchain: a systematic exposition of policies, research hotspots, applications, and prospects. Int. J. Energy Res..
Zhang, C., Benz, P., Karjauv, A., Kweon, S.I., 2021. Data-free universal adversarial perturbation and black-box attack. ICCV 2021.
Zhou, Y., Han, M., Liu, L., He, J., Gao, X., 2019. The adversarial attacks threats on computer vision: A survey. In: 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW). IEEE, pp. 25–30.
Zhu, M., Chen, T., Wang, Z., 2021. Sparse and imperceptible adversarial attack via a homotopy algorithm. ICML 12868–12877.

**Teng Long** is currently an associate professor at School of Information Engineering, China University of Geosciences (Beijing), China. She received her Ph.D. in computer software and theory from State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences. Her research interests include software security, artificial intelligence systems and program analysis.

**Qi Gao** is currently an undergraduate student in the School of Information Engineering, China University of Geosciences (Beijing). His research interests are in the areas of machine Learning and software engineering.

**Lili Xu** is currently a special research assistant at the Institute of Information Engineering, Chinese Academy of Sciences. She received her Ph.D. degree from the University of Chinese Academy of Sciences in 2015. Her research interests include software security, software vulnerability detection and program analysis.

**Zhangbing Zhou** is a professor at China University of Geosciences (Beijing), China, and an adjunct professor at TELECOM SudParis, Evry, France. His research interests include wireless sensor networks, services computing and artificial intelligence.