E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1006-8961(2024)09-2672-20

论文引用格式:Guo K W, Yang K W, Zhang W L, Hu X X and Liu W Z. 2024. A review of adversarial examples for optical character recognition. Journal of Image and Graphics, 29(09):2672-2691(郭凯威, 杨奎武, 张万里, 胡学先, 刘文钊. 2024. 面向文本识别的对抗样本攻击综述. 中国图象图形学报, 29(09):2672-2691)[DOI:10.11834/jig.230412]

面向文本识别的对抗样本攻击综述

郭凯威,杨奎武*,张万里,胡学先,刘文钊

战略支援部队信息工程大学数据与目标工程学院,郑州　450001

摘　要:文本识别技术可以分为光学字符识别(optical character recognition, OCR)和场景文本识别(scene text recognition,STR),其中STR是在OCR基础上针对日益复杂的应用场景衍生出来的。依托深度学习,OCR技术近年来取得了长足进步并大规模商业落地,但深度学习面临的对抗样本攻击问题也给OCR带来了安全威胁。目前大多数OCR模型均存在识别自然扰动和防御对抗样本攻击能力差的问题,如OCR模型在噪声、水印和梯度等攻击算法下的识别准确率大大降低。相比图像领域,文本识别领域的对抗样本攻击研究还远远不够。文本识别通常被视为一个序列到序列的问题,其中输入(如图像中的像素)和输出(像素对应的字符)都是序列,这使得对抗样本的生成更具挑战性。本文对文本识别的对抗样本攻击和防御方法进行研究综述,梳理了近年来文本识别领域的对抗样本攻击方法并进行对比分析,根据攻击类型、应用场景和模型可知性,对攻击方式进行了系统分类。具体来说,按照攻击类型,可分为基于梯度的攻击、基于优化的攻击和基于生成模型的攻击;按照应用场景,可以分为OCR攻击和STR攻击;按照模型可知性,可分为白盒攻击和黑盒攻击。除了回顾文本识别对抗样本攻击方法,还简要介绍了防御技术,具体分为数据预处理、文本篡改检测和传统对抗防御技术。通过这些技术的应用,可以有效地提升文本识别模型的安全性和鲁棒性。最后,总结了文本识别领域对抗样本攻击及防御面临的挑战,并对未来发展方向做出展望。

关键词:光学字符识别(OCR);场景文本识别(STR);对抗样本;生成对抗网络(GAN);深度学习;序列模型

A review of adversarial examples for optical character recognition

Guo Kaiwei, Yang Kuiwu*, Zhang Wanli, Hu Xuexian, Liu Wenzhao

School of Data and Target Engineering, Strategic Support Force Information Engineering University, Zhengzhou　450001, China

Abstract: In the context of deep learning, an increasing number of fields are adopting deep and recurrent neural networks to construct high-performance data-driven models. Text recognition is widely applied in daily life fields, such as autonomous driving, product retrieval, text translation, document recognition, and logistics sorting. The detection and recognition of text from scene images can considerably reduce labor costs, improve work efficiency, and promote the development of information intelligence. Therefore, the research on text detection and recognition technology has practical and scientific value. The field of text recognition has resulted in the use of methods from recognition and sequence networks, which led to evolving technologies, such as methods based on connectionist temporal classification (CTC) loss, those based on attention mechanisms, and end-to-end recognition. CTC- and attention-based approaches perceive the task of matching text images with the corresponding character sequences as a sequence-to-sequence recognition issue by employing an encoder for the process. End-to-end text recognition methods meld text detection and recognition modules into a unified model, which

收稿日期:2023-07-04;**修回日期**:2023-11-23;**预印本日期**:2023-11-30

*通信作者:杨奎武　yangkw@aliyun.com

基金项目:国家自然科学基金项目(62172433,62172434)

Supported by:National Natural Science Foundation of China (62172433, 62172434)

facilitates the simultaneous training of both modules. Although the advancement of deep learning has driven the development of optical character recognition (OCR) technology, some researchers have discovered serious vulnerabilities in deep models: The addition of minute disturbances to images can cause the model to make incorrect judgments. In applications demanding a high performance, this phenomenon greatly hinders the application process of deep models. Therefore, an increasing number of researchers are beginning to focus on strategies for understanding the deep model's response to this anomaly. Before understanding how the model resists this disturbance's performance, a key task is discovering a mechanism for better disturbance generate, i. e. , how to attack the model. Thus, most current research focuses on the development of algorithms that can generate disturbances efficiently. This article reviews and summarizes various adversarial examples of attack methods proposed in the field of text recognition. Approaches to adversarial attacks are divided into three types: gradient-, optimization-, and generative model-based types. These categories are further delineated into white- and black-box techniques, which are contingent upon the level of access to model parameters. In the field of text recognition, prevalent attack strategies involve watermark tactics and cleverly embed disturbances within the watermark. This approach maintains the attack success rate whilst rendering the adversarial image perceptibly natural to the human observer. Common attack methods also include additive and erosion disturbances and minimal pixel attack methods. Generative adversarial network-based attacks have contributed to the research on English and Chinese font-style generation. They deceive machine learning models by producing examples similar to the original data and thereby improve the robustness and reliability of OCR models. The research on Chinese font-style conversion can be attributed to three categories: 1) Stroke trajectory-based Chinese character generation methods generate novel characters by scrutinizing the stroke trajectory inherent to Chinese script. These techniques harness the unique stroke traits of Chinese characters to engender properties with similar stylistic attributes to accomplish style transference; 2) Style-based Chinese character generation methods generate new Chinese characters of specific style by learning the style features of various fonts; 3) Methods based on content and style features generate Chinese characters with specific style and content by learning the representation of content and style features. The attack of OCR adversarial examples provoked reflections on the security of neural networks. Some defense methods include data preprocessing, text tampering detection, and traditional adversarial sample defense methods. Finally, this review summarizes the challenges faced by adversarial sample attacks and defenses in text recognition. In the future, the transition from a white-box environment to a black-box environment requires extreme amount of attraction. In classification models, the content of black-box queries is relatively direct object, with only the unnormalized logical output of the last layer of the model needed to be obtained. However, sequence tasks are incapable of performing single-step output, which makes more effective attacks in fewer query environments a challenging problem. Considerable advancements have been attained in response generation during the advent of substantial vision-language models, such GPT-4. Regardless, the associated privacy and security concerns warrant attention, and thus, the adversarial robustness of large models needs further research. This review aims to provide a comprehensive perspective for the comprehension and resolution of adversarial problems in recognition to find the right balance between practicality and security and promote the continuous progress of the field.

Key words: optical character recognition (OCR); scene text recognition (STR); adversarial examples; generative adversarial network (GAN); deep learning; sequence model

0 引　言

在深度神经网络的推动下，图像分类、目标检测、语音识别和自然语言处理等领域都取得了巨大进步。其中，光学字符识别(optical character recognition，OCR)作为计算机视觉的子领域，在机器翻译、车牌识别、路标识别、垃圾邮件检测、盲人和低视力人群的辅助阅读等方面发挥着重要作用。此外，OCR技术还在场景文字检测、识别方面也发挥着积极的作用。

然而，由于深度神经网络容易受到对抗样本的攻击而造成性能下降，因此当前以深度学习为基础的OCR技术也同样面临对抗样本攻击的威胁。有关对抗样本的生成与防御技术，已有多篇研究论文进行了系统的阐述。张田等人(2022)从特征学习和

分布统计的视角对图像对抗样本的检测与防御进行了深度的探讨；台建玮等人(2022)全面回顾了语音识别系统面临的对抗样本攻击和其防御策略；Ren等人(2021)则系统总结了物理域对抗样本攻击技术；袁珑等人(2022)详述了目标检测系统的对抗攻击方法；杜小虎等人(2021)专门针对自然语言处理中文本情感分析的对抗样本进行了综述。以上论文所提及的对抗样本方法在文本检测等方面有着一定的适用性，但由于OCR是一种与目标检测、情感分析不同的序列分类任务，这些攻击算法无法直接迁移到OCR领域。同时，在OCR对抗样本攻击领域目前尚缺乏系统的综述性文章。

文本识别技术在现代生活中的应用越发广泛，尤其在版权保护、验证码安全以及自动驾驶的道路信息识别等方面。研究文本识别的对抗样本攻击技术，能进一步增强这些应用场景中的数据安全与识别准确性。本文综述了国内外有关文本识别对抗样本的研究现状及成果，首先根据对抗样本的特点，将对抗攻击方法分为基于梯度、基于优化和基于生成的攻击，并对文本识别对抗样本攻击方法进行总结和分析。然后从数据预处理、文本篡改检测和传统对抗样本的防御方法3个方面介绍文本识别的防御技术，最后，对文本识别的对抗样本研究面临的挑战和发展趋势进行展望。

1 概 述

1.1 文本识别技术发展

OCR是指对文本资料的图像文件进行分析识别处理，从而提取文字和版面信息的操作过程。

OCR的概念最初由德国的Tausheck在1929年提出。随后，美国的Handel也提出了类似的文字识别理念，但直到计算机的问世，这些想法才逐步变为现实。1996年，IBM公司的Casey和Nagy采用模板匹配的方法，首次实现了1 000个印刷体汉字的识别。1998年，LeNet-5网络的出现，深度学习开始用于文本识别。直到2012年，AlexNet网络(Krizhevsky等，2017)在ImageNet比赛中取得了超越传统方法的精度，标志着视觉相关技术的重大突破。此后，识别网络和物体检测框架的创新进一步推动了OCR技术的持续发展。

传统OCR技术经历了图像预处理、版面分析、图像切割和特征提取等多个阶段，形成了成熟且完善的技术体系。然而，面对排版多样、背景复杂和分辨率恶劣等挑战，研究者们逐渐放弃了传统的字符切割识别方法。相反，众多基于深度学习的优秀方法应运而生。深度学习的文本识别方法主要将流程拆解为两个核心环节：文本检测和文本识别。前者用于定位文本，后者负责识别文本内容。图1展示了传统OCR和基于深度学习的OCR识别过程。

文本检测领域早期借鉴了物体检测经典的Faster R-CNN(faster region-based convolutional neural network)网络(Ren等，2017)和YOLO(you only look once)网络(Redmon和Farhadi，2018)，发展出了一系列文字专用检测技术，如DenseBox(Huang等，2015)、CTPN(connectionist text proposal network)(Tian等，2016)、TextBoxs系列(Liao等，2017)、PRPN(progressive region prediction network)(Zhong等，2022b)和DBNet++(differentiable binarization network ++)(Liao等，2023)等。在文本识别领域借鉴了识别网络和序列化网络，演变出了多种方法，包括基于连接时序分类(connectionist temporal classification，CTC)损失的方法、基于注意力机制(attention)的方法和端到端识别方法。基于CTC和基于attention的方法将文本图像和对应的字符序列看做是序列到序列的识别问题，通过编码器进行处理。而端到端文本识别方法则是将文本检测模块和文本识别模块集成在一个模型中，以实现同时训练。

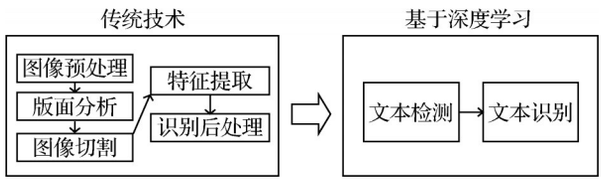

图1 传统OCR和基于深度学习方法的OCR识别过程

Fig. 1 The recognition process of traditional OCR and OCR based on deep learning methods

1)基于CTC的识别方法。如代表性的CRNN(convolutional recurrent neural network)(Shi等，2017)，主要通过卷积神经网络(convolutional neural network，CNN)提取输入图像的卷积特征，利用循环神经网络(recurrent neural network，RNN)建模视觉上下文，最终采用CTC解码识别结果。这种方法通过并行解码和CTC损失提高了解码效率和准确性。

但是，它主要适用于水平文本识别，对于扭曲的文本效果较差，且容易受到无关区域的噪声干扰，因为CTC在每个时间步预测时均独立访问整个图像。后来，一些学者针对CTC网络进行改进。例如，Wan等人(2019)针对自然场景文本的二维特性，在垂直方向增加了一个维度，将CTC网络从一维扩展到二维。这种方法不仅能在缺乏字符级注释的情况下进行训练，还能同时处理字符的分类和位置信息。实验表明，此方法在不规则文本识别上表现优异。

2)基于Attention的识别方法。例如RARE(robust text recognizer with automatic rectification)(Shi等，2016)和ASTER(attentional scene text recognizer)(Shi等，2019)，采用CNN和RNN作为编码器和解码器来预测字符概率。通过Attention机制，在编码器和解码器之间提取特征并预测当前时间步的识别结果。该方法减少无关区域的噪声干扰，提升了识别精度。然而，该方法的串行解码导致效率降低。为应对这一挑战，Luo等人(2021)将生成对抗网络(generative adversarial network，GAN)与基于注意力的识别网络相结合，生成了简化背景的图像，以减少复杂背景的干扰。虽然这些方法具有广泛的适应性，但卷积运算量较大。为了解决这一问题，Zhong等人(2022a)提出了SGBANet(semantic GAN and balanced attention network)，这是一个融合了语义GAN和平衡注意力的网络。该网络通过GAN提取语义特征，确保复杂与简单图像特征的一致性，以提高文本图像的可读性。同时，SGBANet集成了一个平衡注意力模块，利用卷积特征来优化注意力权重，有效地解决了注意力偏移问题。

3)端到端识别方法。例如STN-OCR(spatial Transformer network-ocr)(Bartz等，2017)、FOTS(fast oriented text spotting)(Liu等，2018)、ABCNet(attentive bilateral contextual network)(Li等，2021)、PAN++(pixel aggregation network++)(Wang等，2022)，能同时完成场景文本检测和识别。这些方法能直接实现端到端的识别，无需分步进行检测和识别。端到端识别模型更为轻量、响应更快，适用于高实时性的场景。但其在文本检测与识别特征共享上存在问题，识别精度仍有提升的空间。

1.2 对抗样本攻击

深度学习的进步推动了OCR技术的发展，但同时也催生了众多安全问题，如数据投毒、后门攻击(Li等，2022)、对抗样本攻击。Szegedy等人(2014)研究发现，通过在原始数据中添加精心设计的微小扰动，导致模型输出错误或性能下降。Szegedy等人(2014)认为，对抗样本位于数据分布的低概率区域，即盲区。未覆盖盲区的训练数据生成的模型泛化性能较差，从而导致对抗样本的产生。理论上，同一对象在输入空间的某个邻域内可以从多个角度进行表征，使得邻域内的数据应具有相同的输入标签和统计分布。Goodfellow等人(2015)猜想高维空间下深度神经网络的线性行为是对抗样本存在的原因，提出了图像分类攻击方法的快速梯度符号法(fast gradient sign method，FGSM)。FGSM以其直观的攻击方式和强大的攻击能力，成为对抗攻击领域中最具代表性的方法之一。

许多学者在FGSM的基础上提出了衍生方法。Kurakin等人(2017)将迭代思想引入FGSM，提出了I-FGSM(iterative-FGSM)。Dong等人(2018)在迭代过程中引入动量，提出了MI-FGSM(momentum iterative fast gradient sign method)，可以在迭代过程中稳定更新方向。PGD(project gradient descen)攻击(Madry等，2019)可视为FGSM的多步迭代版，其理念在于：相比于FGSM的单步大幅迭代，PGD选择多步小幅迭代，并在每次迭代后将扰动投射到规定范围内。此外，经典的对抗攻击方法还包括Carlini和Wagner(2017)针对防御蒸馏网络(Hinton等，2015)提出的C&W(Carlini-Wagner)攻击，Su等人(2019)提出的单像素攻击，以及Moosavi-Dezfooli等人(2017)提出的通用对抗扰动(universal adversarial perturbations，UAP)。如图2所示，图中第1行展示了原始图像和不同的攻击版本；而第2行显示了对

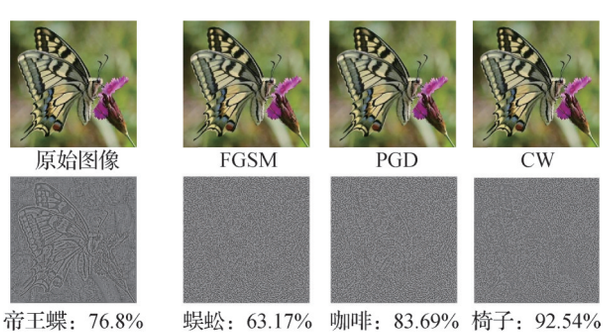

图2 不同攻击方法的结果对比(Kherchouche等，2022)

Fig. 2 Comparison of results from different attack methods (Kherchouche et al. ，2022)

应图像的MSCN(mean subtracted contrast normalized)系数分布图。原始模型以76.8%的置信度将图像判定为“帝王蝶”,而在使用不同攻击方法后,模型分别以63.17%、83.69%和92.54%的置信度将图像错误地判断为蜈蚣、咖啡和椅子。

一些研究者将GAN的生成技术与对抗攻击相结合,例如Baluja和Fischer(2017)提出的对抗变换网络(adversarial transformation networks, ATN),Xiao等人(2018)提出的AdvGAN(adversarial GAN),以及Zhu等人(2023)提出的LIGAA(generative adversarial attack method with low-frequency information),利用低频信息生成对抗样本。对抗样本的发展路径如图3所示。这些研究凸显了网络架构和模型容量在提高对抗鲁棒性方面的重要性。在当前阶段,由于神经网络和数据的高度复杂性,深度学习模型往往具有不可解释性。因此,针对这一问题,不同的假设和研究角度会有不同的侧重点,至今尚未在数理层面达成统一的理解。

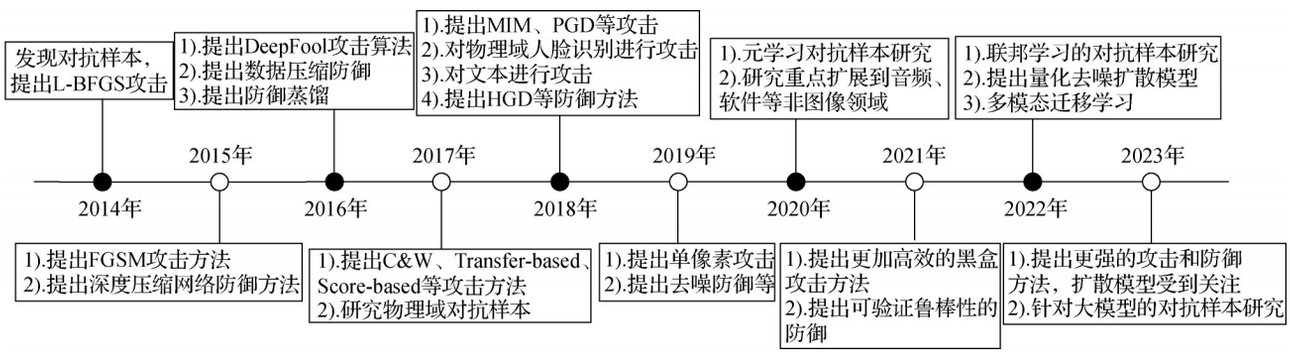

图3 对抗样本的发展路径

Fig. 3 Development path of adversarial samples

同样基于深度学习的OCR模型也受到对抗样本攻击的威胁,因此通过进行关于OCR的对抗样本研究,可以提高OCR模型的鲁棒性。

1.3 评价指标和相关数据集

OCR对抗样本攻击中通常采用的衡量标准包括攻击成功率、扰动平均距离和平均迭代次数。

1)攻击成功率。这是一个定量指标,用于衡量在给定迭代次数下,攻击成功的样本在数据集中的比例。攻击成功率越高,说明攻击算法在单位时间内能生成更多的对抗样本,效果越好。

2)扰动平均距离。表示测试对象中所有攻击成功的图像对应扰动的平均距离。扰动平均距离越小,意味着对抗样本与原图之间的平均差异越小,越难以被肉眼察觉。

3)平均迭代次数。该指标反映了成功攻击一个图像所需的平均迭代次数。次数越少算法越高效。

大多数自然场景文本识别(scene text recognition,STR)算法采用合成数据集进行训练。常用的两个合成数据集包括包含890万幅图像的MJ(MJSynth)和包含550万幅图像的ST(synthtext)。自然场景文本识别的评估数据集主要分为两类:规则文本数据集(如IIIT5K(IIIT5Kwords)、SVT(street view text)、IC03(ICDAR 2003)和IC13)和不规则文本数据集(主要特征为弯曲和透视变换,如IC15、SVT-P和CUTE80)。表1列出了文本识别常用的数据集。这些数据集用于实验和评估各种文本识别对抗样本生成方法,进而推动文本识别技术的发展和完善。

2 面向文本识别的对抗攻击方法

自然语言处理(natural language processing,NLP)领域也应用了对抗样本,如文本分类攻击。类似于图像攻击,在文本分类任务中,通过改变文本中的少量单词或字符,从而欺骗模型产生错误分类。例如,Lei等人(2019)对情感分析和假新闻检测任务进行了文本分类攻击。

现行的大多数攻击策略主要针对具有较大像素值范围的彩色或灰度图像设计,而对像素值范围有限的二值图像(如文本类图像)的效率和适用性不佳。许多OCR模型,如流行的Tesseract(2016)模型,在处理过程中使用二值图像,广泛应用于手写字符识别、车牌号识别和银行支票识别等任务。

针对基于序列模型的文本识别对抗样本攻击,目标是最大限度地引发文本识别错误,同时尽量降

表 1　文本识别常用数据集

Table 1　Common datasets for text recognition

数据集	语言	规模(训练/测试)	数据集特点
MJSynth	英文	890万	每词以不同字体、颜色、大小、角度和背景噪声表现，并配有字符级注释。
SynthText	英文	550万	数据集模拟真实场景，含有各类字体、颜色、大小、排列和视角扭曲的文本，同时配有词级和字符级标注及识别内容，适应文本检测和识别模型需求。
IIIT5KWords	英文	5 000(2 000/3 000)	主要用于词级别的场景文本识别，包括自然场景图像和数字化图像。
Street View Text	英文	350(100/250)	来源于谷歌街景图像，用于街景中的文本识别任务。图像具有多样性和复杂性。
ICDAR2003	英文	509(258/251)	包括水平摄像头捕获的场景文本图像。
ICDAR2013	英文	462(229/233)	用于场景文本识别，是ICDAR2003的升级版，包含更多复杂的自然场景图像。
CUTE80	英文	288	专注于曲线文本识别，图像中的文本具有较强的曲率变换。
ICDAR2015	英文	1 500(1 000/500)	以四边形替代矩形框，收录多方向单词级文本，同时存在模糊、光照变化等挑战性问题，为复杂场景文本识别提供实验环境。
SVT Perspective	英文	639	来自谷歌街景的非正面角度拍照，SVT数据集的扩展，主要关注视角扭曲的文本识别问题。
MSRA-TD500	中英文	500(200/300)	用于测试和评估多方向、多语言文字检测算法的自然图像数据集。
CTW	中文	32 285	来源于腾讯街景。

低对人类视觉的影响。许多现有的图像攻击算法主要针对非序列模型，不适用于序列模型。本文综述了文本识别领域的对抗样本攻击方法，梳理并对比分析了近年来的各种方法，如图4所示，并按照攻击类型、应用场景和模型可知性对其进行了系统分类。具体来说，本文按照攻击类型将其分为基于梯度的

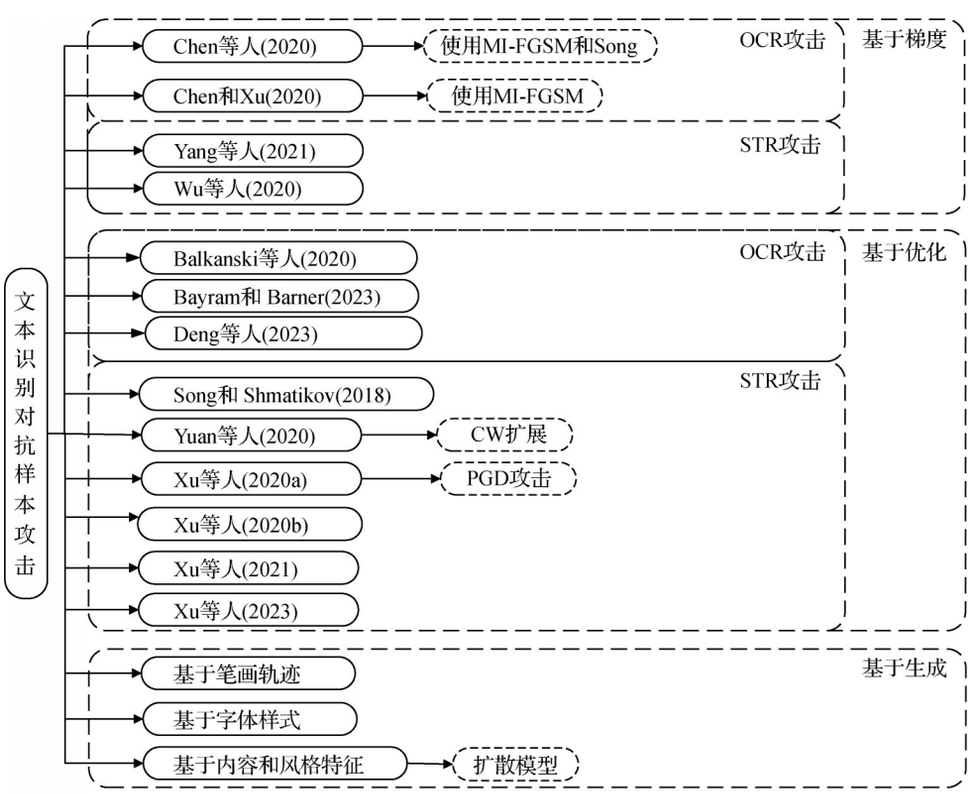

图4　文本识别对抗样本攻击方法分类

Fig. 4　Classification of adversarial example attack methods in text recognition

攻击、基于优化的攻击和基于生成的攻击；按应用场景分为OCR攻击和STR攻击；按模型可知性分为白盒攻击和黑盒攻击。虽然白盒攻击研究较为丰富，但由于对抗样本的迁移性，黑盒攻击（Tramèr等，2017）更具实用价值。

2.1 基于梯度的文本识别攻击

2.1.1 FAWA攻击

Chen等人（2020）提出了一种名为快速对抗水印攻击（fast adversarial watermark attack，FAWA）的白盒攻击方法，巧妙地将扰动伪装在水印中，从而在保持攻击成功率的同时，使得对抗图像在人眼看来显得自然。FAWA使用MI-FGSM和Song和Shmatikov（2018）的攻击方法，将其生成的扰动隐藏于水印中。FAWA评估了Calamari-OCR（Wick等，2018）对于5种不同字体的英文文本，其在单词、语句级攻击中成功率达到100%。

在字母单词水印攻击的基础上，Chen和Xu（2020）针对汉字提出了一种基于梯度的攻击方法。通过不断迭代利用MI-FGSM和CTC损失函数生成扰动，MI-FGSM通过引入动量项，使得每一步的更新不仅考虑当前的梯度，还考虑了之前的梯度信息，既可以稳定更新方向，也可以避免在迭代过程中陷入低效的局部最大值，从而产生具有更优迁移性的对抗样本。为了生成有针对性的L_∞边界对抗样本$\boldsymbol{x}'$，从给定原始图像$\boldsymbol{x}$的初始图像$\boldsymbol{x}'_0=\boldsymbol{x}$开始，MI-FGSM通过求解约束优化问题来寻找对抗样本，具体为

$$\begin{aligned}&\arg\min_{\boldsymbol{x}'} l_{\mathrm{CTC}}(\boldsymbol{x}',t)\\&\text{s.t.}\left\|\boldsymbol{x}'-\boldsymbol{x}\right\|_\infty\leqslant\epsilon,(1-\boldsymbol{\Omega})\odot(\boldsymbol{x}'-\boldsymbol{x})=0\end{aligned}\tag{1}$$

式中，ϵ代表扰动的大小，水印区域用$\boldsymbol{\Omega}$表示，$\boldsymbol{\Omega}=\{o\in\boldsymbol{\Omega}\mid o=1\}$表示在水印内，否则，$o=0$，$\boldsymbol{\Omega}\in\{0,1\}^{|x|}$，$t$表示目标标签，$\odot$符号表示元素逐个相乘。在迭代过程中，首先将获得的对抗样本$\boldsymbol{x}'$送入OCR模型中，并通过反向传播获得梯度，计算识别结果与期望目标结果之间的损失，利用MI-FGSM算法计算出合适的扰动，然后将扰动添加到图像水印区域中，通过不断迭代，直至OCR识别结果达到预期。水印攻击流程如图5所示。

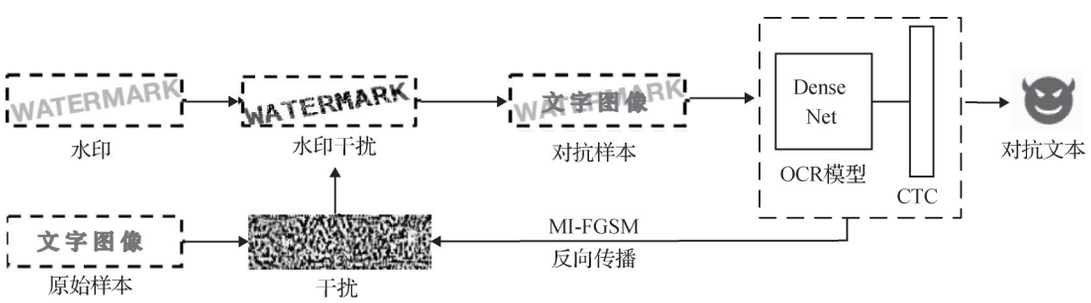

图5 水印攻击流程

Fig. 5 Watermark attack process

为了更好地探究不同水印扰动方法的攻击效果及其呈现的真实水印外观自然度，研究人员设计了WMinit、WMneg和WMedge 3种变体。

WMinit是第1种变体。通过在初始化时添加水印的灰度值来解决生成水印过于稀疏的问题。

$$\boldsymbol{x}'_0=\boldsymbol{x}+\lambda\times(\boldsymbol{x}>\tau)\odot\boldsymbol{\Omega}\tag{2}$$

式中，λ表示水印的灰度，$\boldsymbol{x}>\tau$表示将灰度填充到水印区域中没有文字的位置。

WMneg是第2种变体。其考虑的核心问题是，像素的扰动是基于损失梯度的符号函数生成，当梯度为正，像素值增加，像素呈现更白；反之，像素值降低，像素显得更黑。但对于已经是白色的背景，增加的像素值无效（无法使白色背景更亮），因此，研究者仅保留了对抗样本中为负数的扰动部分。*Clip*函数确保将扰动限制在ϵ的范围内，具体为

$$\boldsymbol{x}'_{i+1}=\boldsymbol{x}'_i+Clip_\epsilon(\alpha\times\boldsymbol{\Omega}\odot\min(0,\mathrm{sign}(\boldsymbol{g}_{i+1})))\tag{3}$$

式中，$\boldsymbol{g}$为动量。

WMedge是第3种变体。通过形态学方法获取文字区域，并将水印扰动区域限制在文本边缘周围，使对抗样本的效果看起来像打印机打印不清晰，使用一个3×3的矩形结构元素作为内核k。具体为

$$\boldsymbol{\Omega}_{\mathrm{edge}}=\{o\in\boldsymbol{\Omega}_{\mathrm{edge}}\mid o=\begin{cases}1 & \boldsymbol{x}\ominus k>\tau\\0 & \boldsymbol{x}\ominus k\leqslant\tau\end{cases}\}\tag{4}$$

式中，$\ominus$符号表示侵蚀操作。

Chen和Xu(2020)的研究为基于序列的OCR模型提出了一种快速对抗水印攻击方法，成功地实现了高攻击成功率和自然的对抗样本外观。通过设计3种变体，进一步探讨了不同水印扰动方式的攻击效果和自然度，为序列学习任务中的对抗攻击研究提供了重要参考。

2.1.2 无目标注意力机制攻击

Yang等人(2021)的研究主要集中在无目标攻击上，即只要对抗样本的预测结果与原始图像产生的预测结果不同，攻击就视为成功。在研究中，采用了基于注意力机制的ASTER作为场景文本识别模型。该模型采用序列到序列学习框架，由编码器和解码器组成。编码器利用深度残差网络ResNet(residual network)提取图像特征，而解码器则通过注意力机制和空间变换器网络(spatial Transformer networks，STN)对特征进行自适应处理并输出最终文本结果。Yang等人(2021)在攻击场景文本识别模型的过程中，提出了一种有效的损失函数，在减少扰动量的同时，实现更高的攻击成功率。具体为

$$L(\boldsymbol{I}+\delta,\boldsymbol{y}')=S\times H\times\sum_{t=1}^{T}\max(d_t,0) \tag{5}$$

$$d_t=\boldsymbol{q}_t^{\boldsymbol{y}_t}-\max_{\boldsymbol{y}_t'\neq\boldsymbol{y}_t}\boldsymbol{q}_t^{\boldsymbol{y}_t'}$$

$$H=\text{sigmoid}(k\times\min_t\{d_t\})$$

$$S=\prod_{t=1}^{T}p(\boldsymbol{y}_t)$$

式中，$\boldsymbol{I}$是输入图像，$\boldsymbol{y}'=(\boldsymbol{y}_1',\cdots,\boldsymbol{y}_{T'}')$是攻击后的目标标签，在步骤$t$中，解码器输出的结果是$\boldsymbol{y}_t$，$\delta$是$\boldsymbol{I}$的对抗扰动，$\boldsymbol{q}$是所有类别上的logit向量。随着识别分数$S$值的增加，攻击的难度也相应增加，损失$L(\cdot)$也相应增大。此外，$H$确保在任何一个字符受到攻击时，损失函数都将等于零。实际应用中，较大的k值使H近似于阶梯函数，从而满足更好的无目标攻击需求。

基于梯度的文本识别攻击主要包括FAWA和无目标注意力机制攻击方法。FAWA攻击采用了Calamari-OCR模型，并使用由文本生成器生成的数据集进行攻击。相比之下，无目标注意力机制攻击则采用了ASTER模型和通用场景文本数据集。FAWA攻击巧妙地把扰动隐藏在水印中，而无目标注意力机制攻击更加灵活，但可能牺牲了一些攻击的精确性。

基于梯度的文本识别攻击的核心是利用模型的梯度信息来生成对抗性样本。通过计算模型的输出对输入的梯度，攻击者可以找到一个方向，使模型的输出偏离真实标签。这种方法的优势在于无需大量计算资源情况下生成高质量的对抗样本。但其局限性也很明显，如模型的内部结构或参数未知时，攻击效果可能会受到限制。

2.2 基于优化的文本识别攻击

基于优化的对抗样本生成方法通过求解优化问题来寻找最佳的通用扰动向量。虽然其核心目标是解决优化问题，但通常采用梯度下降策略，通过小批量样本计算梯度并更新扰动向量，逐步逼近最优解。

2.2.1 针对文档的OCR对抗样本攻击

Song和Shmatikov(2018)通过对扫描文档中最具语义改变潜力的文字添加扰动，成功地使Tesseract输出84.8%的对抗文本。在希拉里·克林顿的电子邮件语料库上实验，通过修改关键信息，如日期、数字和地址，以及将某些特定单词更改为其反义词，使OCR输出的结果与原始文档含义不同。此外，Song和Shmatikov(2018)还进一步证实，OCR系统作为NLP的关键组件，微小的调整在很大程度上影响NLP模型的输出。对抗问题的整个计算式为

$$\min_{x'} c\times L_{\text{CTC}}(f(\boldsymbol{x}'),t')+\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|_2^2 \tag{6}$$

式中，$\boldsymbol{x}'\in[\boldsymbol{x}_{\min},\boldsymbol{x}_{\max}]^p$，$p$是像素数，确保对抗样本是模型$f$的有效输入，$t$为真实序列，$t'$为目标序列，$\boldsymbol{x}'$为对抗图像。作者通过引入变量法将式(6)更改为

$$\min_{\omega} c\times L_{\text{CTC}}\left(f\left(\frac{\boldsymbol{\alpha}\times\tanh(\omega)+\boldsymbol{\beta}}{2}\right),t'\right)+\left\|\frac{\boldsymbol{\alpha}\times\tanh(\omega)+\boldsymbol{\beta}}{2}-\boldsymbol{x}\right\|_2^2 \tag{7}$$

$$\boldsymbol{x}'=\frac{(\boldsymbol{\alpha}\times\tanh(\omega)+\boldsymbol{\beta})}{2}$$

$$\boldsymbol{\alpha}=\frac{(\boldsymbol{x}_{\max}-\boldsymbol{x}_{\min})}{2},\ \boldsymbol{\beta}=\frac{(\boldsymbol{x}_{\max}+\boldsymbol{x}_{\min})}{2} \tag{8}$$

该式添加了一个新变量ω，以便$(\boldsymbol{\alpha}\cdot\tanh(\omega)+\boldsymbol{\beta})/2$在优化期间自动满足框约束。

研究者在生成NLP模型的对抗文本时采用了一个简单的贪心算法。NLP分类器记为h，t'将作为NLP分类器h的输入，使得$h(t')$的预测类别与正确预测$h(t)$不同。为了简化，假设h是一个二元分类器，其中$h(t)$是输入被分类为正确类别的可能性的评分。首先找到文本t中每个单词w的最佳替换词，

然后选择候选替换词集W,使得w与W中每个词之间的编辑距离小于某个阈值τ。编辑距离的限制允许在生成对抗文本图像时使用较小的扰动。其次将所有最佳的单词替换按照它们在分数中引起的变化降序排列,目的是找出对改变NLP模型预测最具影响力的词。最后,贪心算法将t修改为t',通过替换最具影响力的词,直至满足$h(t')$。

2.2.2 基于自适应权重的攻击方法

Cipolla等人(2018)提出了一种基于原则的多任务权重方法,将偶然不确定性和认知不确定性相结合。通过在一个统一的贝叶斯深度学习框架中为每个任务建模,在语义分割、实例分割和深度回归等方面表现优于单独训练的模型。然而,这种方法主要适用于非序列学习任务(如图像分类、图像分割),并不能直接应用于序列学习任务的对抗攻击。

Yuan等人(2020)将对抗攻击视为一种多任务学习,并将Cipolla等人(2018)提出的多任务权重计算方法扩展到序列学习中,特别是场景文本识别任务中。Yuan等人(2020)通过利用每个任务的不确定性,直接学习自适应多任务权重,无需手动搜索超参数。具体为

$$-\log \Pr(l'|\boldsymbol{x}') \approx -\log(\Pr(\pi_1|\boldsymbol{x}') + \Pr(\pi_2|\boldsymbol{x}')) \leqslant -\frac{1}{2(\log \Pr(\pi_1|\boldsymbol{x}') + \log \Pr(\pi_2|\boldsymbol{x}'))} - \log 2 \tag{9}$$

式中,π_1和π_2为推理过程中的两条路径。

联合损失函数L满足

$$L \leqslant \frac{L_1(\boldsymbol{x}, \boldsymbol{x}')}{2\lambda_1^2} + \frac{l_{\text{CTC}}(\boldsymbol{x}', \boldsymbol{l}')}{n\lambda_2^2} + \log \lambda_1^2 + T\log \lambda_2^2 + \frac{\log n - (n-1)\log c}{n\lambda_2^2} \tag{10}$$

在实验中,CTC损失总是以非常快的速度减少。因此,使用少量的有效路径来生成对抗样本。最小化L的上界来生成序列对抗样本。具体为

$$\frac{L_1(\boldsymbol{x}, \boldsymbol{x}')}{\lambda_1^2} + \frac{l_{\text{CTC}}(\boldsymbol{x}', \boldsymbol{l}')}{\lambda_2^2} + \log \lambda_1^2 + T\log \lambda_2^2 + \frac{1}{\lambda_2^2} \tag{11}$$

式中,$\boldsymbol{x}'$为对抗样本图像,λ_1为噪声等级,λ_2为正标量,真实序列标签$\boldsymbol{l} = \{l_0, l_1, \cdots, l_T\}$,目标序列标签$\boldsymbol{l}' = \{l_0', l_1', \cdots, l_T'\}$ $(\boldsymbol{l} \neq \boldsymbol{l}')$。

实验结果表明,自适应攻击在对场景文本数据集上的成功率超过99.9%,并在合成的MNIST(modified national institute of standards and technology)数据集上成功实现对抗攻击。

2.2.3 场景文本识别对抗优化攻击

Xu等人(2020a)针对场景文本识别模型制作对抗样本时,正式引入问题定义,并将其形式化为一个优化问题。具体为

$$\min_{x'} L(\boldsymbol{x}', \boldsymbol{l}') + \lambda D(\boldsymbol{x}, \boldsymbol{x}')$$
$$\text{s.t. } \boldsymbol{x}' = \boldsymbol{x} + \delta, \boldsymbol{x}' \in [0,1]^{|n|} \tag{12}$$

对于其他模型方面的约束:$D(x, x') = \|\boldsymbol{\delta}\|_2^2$是原图与对抗样本之间的$L_2$范数距离,$L(\cdot)$是攻击的代价函数,针对不同的场景文本识别模型具有细微差异。λ是用于平衡$L(\cdot)$和$D(\cdot)$的超参数,其不同的值会影响最终的结果。从经验角度看,较小的λ可以更容易且更快速地导致攻击成功。

CTC模型的目标式白盒攻击的代价函数可以简化为

$$\min_{\boldsymbol{x}'} ((-\log \sum_{\pi' \in S(l')} \max p(\pi'|\boldsymbol{x}')) + \lambda D(\boldsymbol{x}, \boldsymbol{x}')) \tag{13}$$

CTC场景文本识别模型定义的无目标式白盒攻击代价函数可定义为

$$\min_{\boldsymbol{x}'} ((-\log \sum_{\pi' \in S^{-1}(l')} \max p(\pi'|\boldsymbol{x}')) + \lambda D(\boldsymbol{x}, \boldsymbol{x}')) \tag{14}$$

attention-based模型的目标式白盒对抗攻击算法的最终优化目标是

$$\min(\max(-\log p(l_t'|\boldsymbol{x}, l_1', \cdots, l_{t1}')) + \lambda D(\boldsymbol{x}, \boldsymbol{x}')) \tag{15}$$

attention-based模型的无目标式白盒对抗攻击算法的最终优化目标是

$$\min(\max(-\log p(l_t'|\boldsymbol{x}) + \lambda D(\boldsymbol{x}, \boldsymbol{x}')) \tag{16}$$

式中,$p(\pi'|\boldsymbol{x}')$表示在目标序列上有效的解码路径π'的概率。在最优化问题和代价函数的分析中,仅扰动δ是变量,且参数简单。由于网络的高度非线性复杂性,直接找到全局最优的扰动变得困难。然而,在已知模型结构和参数的白盒攻击场景下,使用随机梯度下降法找到最优扰动δ,并将其加到原图构造对抗样本,该方法本质上是PGD攻击的迁移应用。

2.2.4 基于像素翻转的对抗样本攻击

Balkanski等人(2020)针对OCR系统,提出了Scar的黑盒攻击方法,其原理类似加性扰动,通过在字符背景添加扰动,试图将扰动隐藏在字符附近。但该方法需要较大的扰动来误导分类器。

为了降低查询次数,Scar算法优先评估预计在空间和时间上具有较大增益的像素。若其中一个像

素具有较大的增益，则翻转该像素，而无需评估其他像素。若所有像素增益均不大，将考虑图像文字中黑白区域边界上的所有像素。在这个集合中，无论其增益是否大于阈值，都会翻转具有最大增益的像素。研究者进一步扩展了Scar算法，设计了针对先进的支票处理系统的攻击方法，通过利用不易被察觉的扰动来欺骗系统，从而导致存款金额的错误分类。

在Scar方法的基础上，Bayram和Barner(2023)提出了EcoBA(efficient combinatorial black-box adversarial attack)的黑盒攻击方法。ECoBA方法主要由两部分组成：加性扰动和侵蚀性扰动。这两种方法可以完全控制是否将扰动应用于文字。加性扰动发生在文字的背景中；而侵蚀扰动则出现在文字上。加性扰动表示为

$$\arg\min_i F(\boldsymbol{x} + w_i),\ \|\boldsymbol{x}' - \boldsymbol{x}\|_0 \leqslant k \tag{17}$$

设$\boldsymbol{x}$是一个光栅化二进制图像，其中$\boldsymbol{x}$的每个元素均为0(黑色)或1(白色)。为生成附加扰动，逐个翻转背景(黑色)像素，扫描整个图像。记录每个潜在像素翻转对分类器性能的影响，并将结果排序保存在字典中，附带有导致错误的相应像素索引，记为$\boldsymbol{D}_{\mathrm{AP}}$。将从黑色切换到白色的像素表示为$w_i$，其中$i$表示像素索引。重复该过程，$k$表示翻转像素的数量。其中对抗样本$\boldsymbol{x}' = \boldsymbol{x} + w_i$，如果$\epsilon_i = \boldsymbol{x}_i - \boldsymbol{x}_i' > 0$，则保存第$i$个像素索引，否则，重复该过程，跳到下一个像素。侵蚀扰动表示为

$$\arg\min_i F(\boldsymbol{x} + b_i),\ \|\boldsymbol{x}' - \boldsymbol{x}\|_0 \leqslant k \tag{18}$$

侵蚀性扰动是加性扰动的镜像过程。即识别并翻转文字上的白色像素，导致最显著的对抗错误，将从白色翻转到黑色的像素表示为b_i。优化过程确定导致置信度显著下降的像素并翻转。类似地，排序的错误被保存在字典中，记为$\boldsymbol{D}_{\mathrm{EP}}$。

ECoBA可视为加性和侵蚀扰动的综合应用。将$\boldsymbol{D}_{\mathrm{AP}}$和$\boldsymbol{D}_{\mathrm{EP}}$合并到复合字典中，记为$\boldsymbol{D}_{\mathrm{AEP}}$。例如，$\boldsymbol{D}_{\mathrm{AEP}}$的第1行包含最高的$\epsilon$对应$w_i$和$b_i$的值。对于$k=1$，两个像素被翻转，分别对应于$w_1$和$b_1$，导致黑色(或白色)像素数量没有复合变化。也就是说，L_0范数没有变化。因此，在使用k作为迭代索引的过程中，k对应于翻转像素对的数量和扰动的数量。

2.2.5　基于自适应差分进化的黑盒攻击

Xu等人(2021)提出了一种基于差分进化算法(differential evolution algorithm，DEA)的黑盒攻击方法，通过最小化像素干扰来攻击场景文本识别模型。DEA作为一种随机并行直接搜索算法，对优化目标几乎无需假设，并且不需要梯度信息，能在高维空间搜索解，通过迭代寻找优化问题的最优解。但DEA需要大量查询，可能导致在目标模型上的查询次数过多。为了改进DEA，有研究者提出了动态差分进化(dynamic differential evolution，DDE)，它具有内存需求更低且查询次数更少的特点。DEA通过逐代更新对抗扰动，而DDE通过逐个体更新对抗扰动，能够加速收敛，但容易陷入局部最优，且通常耗时较长。Xu等人(2023)结合DEA和DDE优势，提出自适应差分进化(AD^2E)方法，以平衡性能、速度和查询次数。

AD^2E包括适应、突变、交叉及选择4个阶段。在适应阶段旨在估计当前一代是否具有“期望”的对抗性扰动。当适应性阶段估计出当前一代中可能已经有接近目标解的最优对抗扰动时，这个最优对抗扰动就会被当做所期望的对抗扰动。随后，选择阶段将自动逐个体地更新当前一代，以加速收敛速度。然而，如果适应性阶段估计出当前一代中不存在所期望的对抗扰动，选择阶段将并行地逐代更新，从而缩短运行时间。在优化过程中，如果适应性阶段发现搜索过程已陷入局部最优解，那么将直接跳过当前的优化过程以减少运行时间。突变和交叉阶段旨在探索搜索空间，而选择过程确保有潜力的探测对抗图像可以进一步生成。

2.2.6　通用防御底部绘制补丁

Deng等人(2023)提出一种通用防御底部绘制补丁(universal defensive underpainting patch，UDUP)的攻击策略，这种机制不是直接操作字符，而是修改文本图像的底层绘制。通过迭代优化方法，设计了一个固定大小的防御补丁，可以为各种尺寸的文本图像产生不重叠的底部绘制。这个策略的独特之处在于，它并不尝试对被攻击的文本进行识别，而是利用当前流行的文本检测框架进行效果评估，使这些框架无法准确标记出文字区域。这种攻击方法与字符的具体内容、尺寸、颜色和语言均无关，而且对于常见的图像缩放和压缩操作都表现出了良好的鲁棒性。

如表2所示，Xu等人(2023)对文本识别的主要攻击方法在IC13、IC15、IIIT5K、CUTE数据集上进行了定量比较，其中SR(success rate)代表攻击成功率，

L_0代表原始图像与对抗图像之间的L_0距离。结果表明，在模型参数和像素较少的情况下，AD^{2}E仍能实现接近的SR。基于优化的文本识别攻击，相较于基于梯度的方法，展现出更高的攻击效率，特别是在黑盒攻击场景中。它不仅能够快速地生成对抗样本，而且可以进行有目标式的攻击，使得模型输出特定的错误结果。这种方法的灵活高效使其在文本识别攻击中具有明显的优势。

表2　文本识别主要攻击方法定量比较

Table 2　Quantitative comparison of main attack methods in text recognition

方法	目标模型	威胁模型	IC13		IC15		IIIT5K		CUTE	
			L_0	SR	L_0	SR	L_0	SR	L_0	SR
Yang等人(2021)	ASTER	白盒	–	**100**	–	**100**	–	**100**	–	**100**
Xu等人(2020a)	CRNN	白盒	207 6	99.5	113 4	99.9	184 5	99.9	205 4	**100**
Xu等人(2020a)	Rosetta	白盒	189 9	99.7	113 1	**100**	170 9	**100**	213 5	**100**
Xu等人(2020a)	STAR	白盒	217 6	96.2	229 0	99.7	191 6	98.1	229 0	**100**
Xu等人(2020a)	TRBA	白盒	224 4	93.5	234 6	99.6	194 1	97.1	234 6	99.1
Xu等人(2021)	CRNN	黑盒	7	98.0	7	99.4	**7**	98.3	6	99.5
Xu等人(2021)	Rosetta	黑盒	6	99.6	7	100	**7**	98.8	6	**100**
Xu等人(2021)	STAR	黑盒	7	97.7	7	97.6	**7**	96.7	**4**	98.5
Xu等人(2021)	TRBA	黑盒	7	95.7	7	97.2	**7**	94.7	7	97.2
Xu等人(2023)	CRNN	黑盒	**5**	99.7	**5**	**100**	**7**	99.8	7	**100**
Xu等人(2023)	Rosetta	黑盒	**5**	**100**	7	**100**	**7**	99.7	**4**	**100**
Xu等人(2023)	STAR	黑盒	6	98.5	7	99.0	**7**	98.1	6	99.5
Xu等人(2023)	TRBA	黑盒	6	97.5	7	98.5	**7**	97.2	7	**100**

注：加粗字体表示攻击成功率的最优结果，“–”表示无相应实验数据。

2.3　基于生成的文本识别攻击

生成对抗网络(GAN)在英汉字体风格生成研究中发挥了至关重要的作用，它通过产生类似于原始数据的样本来欺骗机器学习模型，从而提高OCR模型的鲁棒性。生成对抗网络由生成器(generator)和判别器(discriminator)两部分组成。生成器接收先验随机噪声作为输入，生成对抗图像，通过对抗训练过程使生成图像的分布逐渐接近真实图像分布，达到欺骗判别器的目的。在理想情况下，判别器和生成器将达到零和博弈的状态。通过在生成对抗网络的生成器和判别器中添加额外信息作为条件，可以将其扩展为条件生成对抗网络，以控制网络输出。这些附加信息可以是任何类型的辅助信息，如图6所示。

在条件生成对抗网络中，先验随机噪声向量与附加信息以联合形成条件映射关系。因此，生成器与判别器之间的对抗训练过程可以用以下损失函数表示为

$$L_{\mathrm{CGAN}}(G,D)=E_{\boldsymbol{x}\sim P_r(\boldsymbol{x})}[\log D(\boldsymbol{x}|c)]+E_{z\sim P_z(z)}[\log(1-D(G(z|c)))] \tag{19}$$

式中，$\boldsymbol{x}$代表原始图像，G表示生成器，D表示判别器。$P_r(\boldsymbol{x})$表示真实数据；$P_z(z)$表示随机噪声数据；E_x表示下标分布的期望值；c表示附加信息；生成器和判别器交替地最小化和最大化这个损失函数，最终求得近似最优解的生成式模型G^*。

由于汉字的数量众多且结构复杂，汉字自动生成任务与其他文字生成任务存在明显差异，使得英文的文本生成模型难以直接应用于汉字生成。近年来，汉字字体风格转换的研究主要可归为3类：

1)基于笔画轨迹的方法。利用汉字的笔画特性生成具有相似风格的汉字，实现风格迁移。

2)基于样式的汉字生成方法。通过学习不同字体的样式特征，生成具有特定样式的新汉字。

3)基于内容特征和风格特征的方法。通过学习

内容特征和风格特征的表示，这类方法可以生成具有特定风格和内容的汉字。

利用GAN生成多种汉字风格，可以降低OCR识别率，达到攻击效果。通过将生成的汉字制作成字体库，在输入时将文档中的汉字替换为字库中的字，以防止在资料泄露时敌人批量识别获取有价值信息。此外，将生成的不同风格字体进行对抗训练，又可提升OCR的性能。

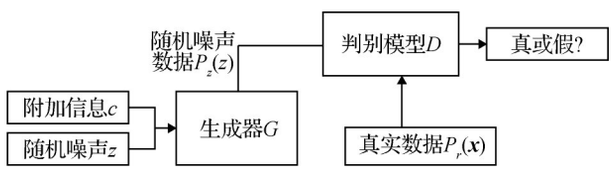

图6　CGAN原理

Fig. 6　Principle of CGAN

2.3.1　基于笔画轨迹的汉字生成方法

为了实现汉字字体风格迁移，研究人员尝试了多种方法。Feng等人(2019) 建立了包含所有汉字偏旁、部首和字形结构的数据表。通过拆分识别出的汉字，查找对应的偏旁部首，进行转换和组合，可以获得改变风格的汉字。但此方法存在一些问题，如转换风格固定、合并后的汉字可能与原字体大小有差异等。

Xu等人(2009)提出了形状语法方法，以多层次方式表示每个字符，并通过基于约束的推理系统生成书法。Zhou等人(2011)提出了汉字部首组合模型，StrokeBank通过将标准字体组件映射到手写对应物来合成字符。Lin等人(2015)设计了一种算法，利用人类标记的位置和大小信息，组装提取的组件来合成目标样式中的给定字符。但这些方法仍然需要专业知识来手动定义基本的建设性元素，并且人的监督和微调在获取完美的笔画提取结果时是不可避免的。Lian等人(2016)开发了一个系统来分别学习笔画形状和布局，避免人为干预。然而，构图规则是基于对参考字体样式的强假设，对于与参考数据相比具有巨大形状差异的手写字体不太适用。

2.3.2　基于样式的汉字生成方法

由于笔画轨迹效果不理想，许多研究人员转向使用风格迁移改编的模型来合成汉字。例如，Tian(2016)提出的Rewrite可以将给定字符从标准字体样式转换为目标样式。zi2zi模型，如图7所示，一个基于pix2pix模型的开源项目，结合了U-Net架构和对抗训练，使用AC-GAN(auxiliary classifier GAN)(Odena等，2017)的多类别损失解决生成字符的风格与目标风格不一致的问题。但是，这些方法无法生成具有复杂形状的高质量图像。

在GAN的帮助下，可以合成更逼真和更高质量的角色图像。Jiang等人(2017)提出的DCFont(deep Chinese font)包含两个主要组件，分别是字体特征重建网络和字体样式传输网络。给定用户书写的少量字符，字体特征重构网络试图估计所有其他字符的深字体特征，字体样式传递网络使用重构的特征将参考字体样式中的字符转换成相应的手写样式。尽管许多学者提出了基于CNN的字体生成网络以模拟特定作家风格，但其实际应用效果并不好。

为了解决上述问题，一些研究人员提出了新的汉字生成方法。Jiang等人(2019)提出了SCFont(structure-guided Chinese font)，通过调整笔画提取算法进一步改进了DCFont，以保持从输入图像到输

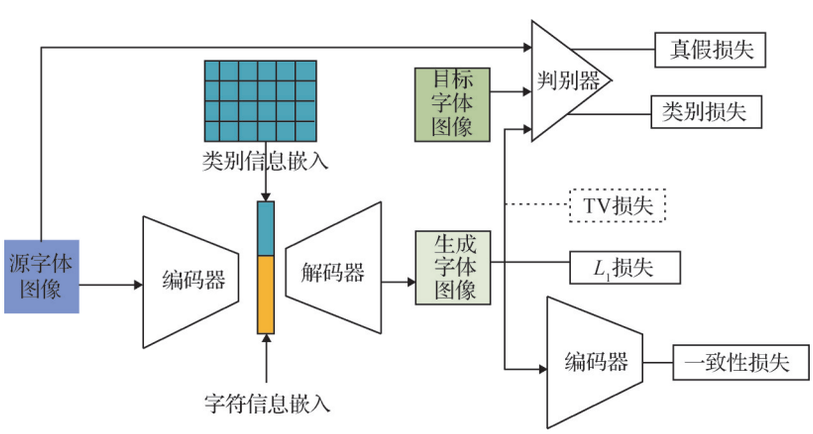

图7　zi2zi模型

Fig. 7　The zi2zi model

出图像的笔画结构。Wang等人(2022)提出的Attribute2Font根据指定字体属性的给定值生成新字体样式的字形图像,同时使用属性注意力模块合成高度多样化且具有美观性的字符图像。

虽然基于样式的汉字生成方法在字符结构上表现出相当的正确性,但在细节的笔画方面仍有不足,例如局部扭曲或遗漏的笔画、笔画的粗细变化、突然和消失的笔锋等问题。对于连笔字体,学习的难度更为增大,这可能源于网络模型在处理大规模字形变化时,提取出适当的特征表示的能力有限。

2.3.3 基于内容特征和风格特征的方法

Wu等人(2020)提出的CalliGAN,通过分解字符为组件并利用额外的网络提取内容信息,实现了对汉字整体风格和结构特征的兼顾。

Xie等人(2021)提出了DG-Font(deformable generative networks for unsupervised font generation),用于无监督字体生成的可变形生成网络,将内容和样式分离为两个不相关的域,并从字体图像中学习潜在样式。不同字体同一个字的笔画具有对应关系,可视为由变形而来,因此提出FDSC(feature deformation skip connection)模块,利用可变形卷积对内容分支进行特征抽取,并结合风格分支的特征进行目标字符生成。Yuan等人(2022)提出的SE-GAN(skeleton enhanced GAN)通过设计自关注的精炼注意力模块来引导模型学习字体的骨骼特征和风格特征,常见的书法字体生成样式如图8所示。虽然这些方法提高了字形生成的质量,但仍存在挑战。例如草书风格的研究较少,其复杂的笔画和结构难以被网络模拟。此外,书法字体由于其独特风格而降低了OCR的识别准确率。

GAN字体生成,从图像风格转移的问题演变而来,将源文本风格转移到目标文本字形,形成具有艺术风格的新文本图像。Yang等人(2019)用任意的文本效果对文本进行风格化,并用参数化的方式控制字形变形的程度。Liu等人(2022)在此基础上提出了具有艺术文本风格的对抗样本,欺骗了STR模型,如图9所示,干扰后的146535识别成521393,$a\rightarrow w$,$e\rightarrow h$,$i\rightarrow f$,$d\rightarrow a$,$t\rightarrow c$。

虽然大多数现有方法依赖于GAN,但它们常常面临对抗损失难以收敛和模式坍塌的问题。广泛使用的图像到图像转换容易忽视跨字体之间的潜在变化。为了克服这些挑战,He等人(2023)提出了Diff-Font,该模型基于多属性条件的扩散模型来进行字体生成,而不是传统的图像到图像转换方法。在这

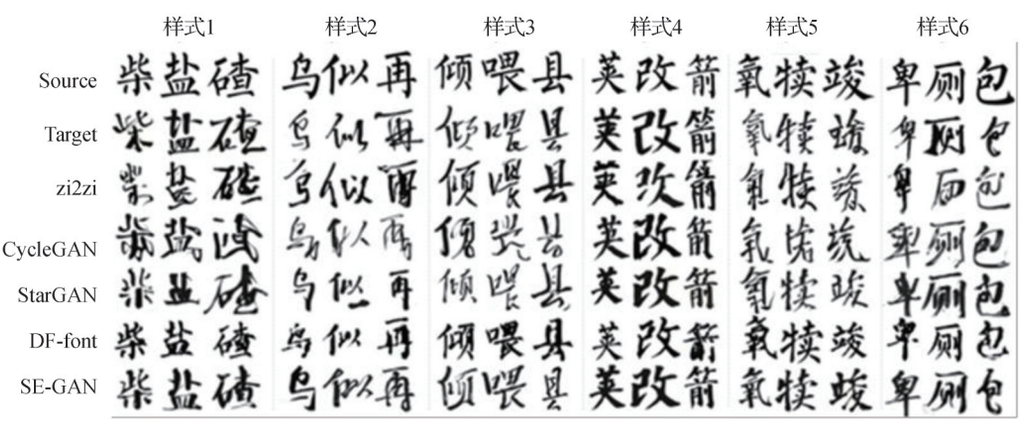

图8 书法字体生成样式(Yuan等,2022)

Fig. 8 Calligraphy font generation style(Yuan et al., 2022)

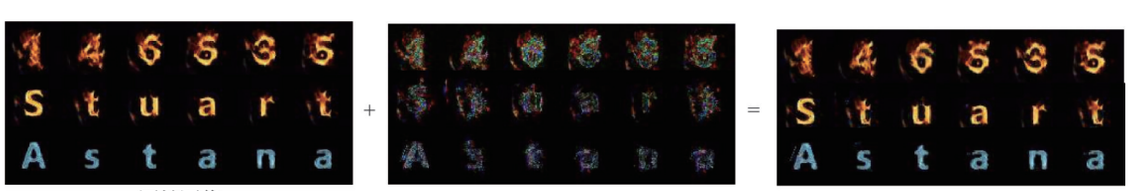

图9 原始和对抗文本图像示例

Fig. 9 Examples of original and adversarial text images

个模型中，字符的内容和风格都被视为生成条件，使得扩散模型能够根据这些条件生成对应的字符图像。为了优化复杂结构的脚本生成，还引入了笔画和组件的细致条件。

然而，与大部分扩散模型一样，Diff-Font也存在推理效率低的问题。虽然引入笔画或组件条件可减少错误，但某些罕见、复杂或独特的字符仍可能生成失败。为了实现高质量的汉字生成，未来的研究需继续探索更有效的方法来克服现有技术的局限性，特别是在书法字体风格转换和OCR识别方面。

基于梯度和基于优化的攻击则是对图像上的文字进行扰动，而基于生成模型则是从生成新字体来解决这个问题，该部分介绍了基于生成模型的文本识别攻击方法。此方法不仅能利用生成模型生成对抗样本以欺骗文本识别系统，还可用于创造新的字体样式，展现了字体生成的重要商业价值和应用潜力。

2.4 OCR对抗样本攻击方法分析

在实际应用中，虽然有些方法能够在受控环境下成功欺骗OCR系统，但在版权保护等应用场景中却面临着很多问题。这些方法通常会为每幅文本图像单独生成对抗扰动，从而导致了它们的通用性不强和效率低。同时，有些方法并未考虑到图像缩放操作对结果的影响。

文本识别对抗样本生成方法多以白盒攻击为主，如表3所示。但在实际中，白盒攻击难以实现，因为攻击者往往无法获取目标模型的详细信息。预期未来研究将更多地转向实用性更强的黑盒攻击，与此同时，OCR对抗样本的主要攻击范式是L_2范数，其主要目的是量化和最小化扰动的大小。未来的研

表3 文本识别对抗攻击方法的全面比较

Table 3 Comprehensive comparison of adversarial attack methods for text recognition

方法	攻击原理	识别模型	威胁模型		攻击类型		扰动限制	攻击场景	优点	缺点
			白盒	黑盒	目标	无目标				
Song和Shmatikov（2018）	基于优化	CTC	√	×	√	×	L_2	OCR NLP	使用贪心算法	数据集单一
Yuan等人（2020）	基于优化	CTC	√	×	√	×	L_2	STR	将多任务权重计算扩展到序列模型	比较耗时
Xu等人（2020a）	基于优化	CTC attention	√	×	√	√	L_2	STR	针对常见的场景文本数据集实验	扰动范围大
Xu等人（2020b）	基于优化	CTC attention	√	×	√	√	L_2	STR	扩展到了图像字幕攻击	仅针对英文实验
Chen等人（2020）	基于梯度	CTC	√	×	√	×	L_2	OCR	水印攻击、针对英文	水印形状位置固定
Chen和Xu（2020）	基于梯度	CTC	√	×	√	×	L_∞	OCR	水印攻击、针对中英文	水印形状位置固定
Yang等人（2021）	基于梯度	attention	√	×	×	√	L_2	STR	成功率高	无法目标攻击
Xu等人（2023）	基于优化	CTC attention	×	√	×	√	L_0	STR	自适应差分进化算法	图像尺寸有要求
Balkanski等人（2020）	基于优化	LSTM	×	√	×	√	L_0	OCR	加性扰动	扰动明显数据集单一
Bayram和Barner（2023）	基于优化	MLP2LENET	×	√	×	√	L_0	OCR	加性和侵蚀扰动组合	数据集单一
Deng等人（2023）	基于优化	PAN++ EasyOCR	√	√	√	×	–	OCR	基于扩散模型、图像大小无限制	推理效率低
Liu等人（2022）	基于GAN	CTC attention	√	×	√	×	L_∞	STR	艺术风格字体生成	针对英文实验

注：“√”表示采用该模型；“×”表示未采用该模型，“–”表示原文未说明。

究可能会探索其他扰动范数，以在不同场景下平衡攻击效果与扰动大小。

总体而言，对抗样本研究包括了多种攻击策略，例如基于梯度的攻击、优化攻击以及结合生成模型的攻击等。这些策略共同揭示了深度学习模型在对抗鲁棒性方面的弱点，为未来深度学习模型的鲁棒性研究提供方向。

3 文本识别的防御技术

OCR对抗样本的攻击，引发了人们对神经网络安全性的反思，进而提出了一些防御方法，如图10所示，针对文本识别对抗样本的防御策略主要包括数据预处理、文本篡改检测与传统的对抗样本防御方法。

3.1 数据预处理方法

OCR对抗样本本质上是添加噪声。常见的防御方法是通过降噪消除干扰，提高准确率。常见降噪方法包括图像裁剪或缩放、位深度缩减、图像压缩和总方差最小化等。降噪的目的是通过转换对抗样本，破坏噪声干扰，消除扰动影响，从而提高准确率。根据神经网络结构，在输入层和隐藏层进行降噪。然而，现有方法依赖梯度遮掩，这一问题并未得到很好解决，导致目前使用降噪技术的防御方法在BPDA（backward pass differentiable approximation）（Athalye等，2018）等对抗攻击下表现出脆弱性。

Prakash等人（2018）通过像素偏转算法重新分布像素值破坏输入图像，利用小波域中的自适应软阈值使模型输出平滑来抵御对抗扰动。在Mustafa等人（2020）的研究中，图像恢复过程使用小波去噪和超分辨率技术。然而，与任何图像去噪算法一样，该技术在消除干扰的同时，会将其他失真添加到图像内容中。Guo等人（2018）证明了在将图像输入神经网络前进行预处理（位深度缩减、总方差最小化等）也可减少对抗噪声带来的影响。

另一种防止对抗攻击的方法是特征去噪，在推理过程中，在某些模型层之后对输入样本的特征进行去噪处理。Xie等人（2019）使用非局部均值滤波器对图层输出进行去噪。Borkar等人（2020）引入了选择性特征再生作为去噪过程。Liao等人（2018）提出了一种高级表示指导去噪器（high-level representation guided denoiser，HGD），使用U-Net作为去噪网络，通过定义损失函数为干净图像和去噪图像激活的目标模型输出之间的差，解决了残余噪声放大问题。Jia等人（2019）设计了ComDefend，通过利用图像压缩来消除或破坏对抗扰动。ComDefend针对清晰图像进行网络训练，学习清晰图像的分布，从而能够从对抗图像中重构清晰图像。与HGD和PixelDefend不同的是，ComDefend在训练阶段无需对抗样本，从而降低了计算的成本。ComDefend采取的是对图像进行逐块处理，而非直接处理整个图像，提高了处理效率。

此外，本文在前面内容中提到了水印攻击，可以使用水印去除技术，使用OpenCV中实现的修复方法（Telea，2004），该方法需要水印的掩码作为先验，并尝试根据周围像素恢复水印区域。在修复去除水印的同时，由于文本区域与水印区域重叠，修复方法去除了太多有用的像素，导致OCR完全失败，甚至影响了文本的可读性。

然而，这些方法并不能完全消除噪声带来的影

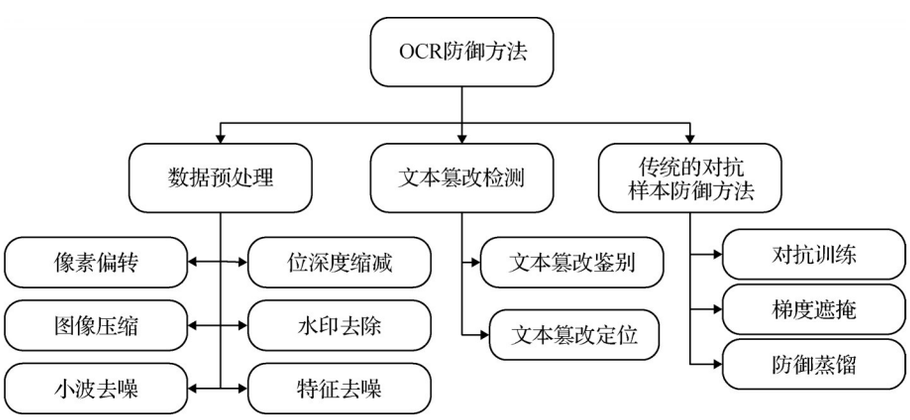

图10 OCR防御方法

Fig. 10 OCR defense methods

响，使得它们在白盒攻击下缺乏鲁棒性。通过对OCR训练的图像数据预处理，可实现数据增强的目的。

3.2 文本篡改检测

图像中的文字携带着丰富的信息，特别是在金融和行政等领域，这为伪造和篡改提供了机会。实际上，句子中的微小变化可能显著改变其携带的语义信息。然而，过去的文档分析与识别研究主要关注检测和理解文本内容。近年来，图像取证受到学术界和工业界越来越多的关注（李晓龙 等，2021），目的是防范恶意图像操纵。大部分研究集中在自然图像上，其中被篡改的主体通常是人或汽车等物体。由于文本的非结构化呈现，篡改文本检测面临更大的挑战性。例如，篡改区域可能非常小（如段落中的一个字符）、篡改区域与周围环境之间的对比度可能非常低。

文本篡改检测可分为两个部分：文本篡改鉴别和篡改文本定位。文本篡改鉴别涉及模型判断输入的文本图像是否包含篡改文本，并输出分类结果。文本篡改定位则要求模型进一步确定篡改文本的具体位置。

Joren等人（2022）将拼接检测问题视为图节点分类问题，利用OCR技术获取文档图像中文本块的位置和内容，基于此结果构建图神经网络。具体来说，该方法以各文本块为图节点，节点间的连接由检测框距离决定。基于预训练的变分自编码器对文档图像进行特征提取，进而引入图注意力机制捕获更强的文本块上下文特征，以提高分类准确度。

3.3 传统对抗样本防御方法

常见的对抗样本防御方法有对抗训练，防御蒸馏、梯度遮掩等。对抗训练最初由Goodfellow等人（2015）提出，是最早的防御技术，其核心思想是将生成的对抗样本纳入训练集中，作为一种数据增强手段，使模型在训练时预先学习和适应对抗样本。通过对抗训练提升OCR的性能，也是目前认为最有效防御白盒攻击的方式。然而，对抗训练依赖于训练数据集，可能导致过拟合问题，而知识迁移可以被采用以提升模型的泛化能力。为了增强神经网络模型并抵御攻击，对抗训练方法在训练过程中会将生成的对抗样本加入其中，例如，Goodfellow等人（2015）将使用FGSM攻击生成的样本添加到MNIST数据集的训练过程中，而Madry等人（2019）利用PGD攻击生成的对抗样本进行类似的操作。这些模型对迁移的扰动不具有鲁棒性，因此，Tramèr等人（2020）引入了包含迁移扰动的整体对抗训练。对抗性训练的主要局限性包括：耗费时间，对新的或未知的攻击缺乏鲁棒性，以及对黑盒攻击的健壮性不足。

4 未来研究方向

4.1 目前存在的问题

在前文中已经全面分析了近年来OCR对抗样本攻击与防御的相关研究，从中可以看出，面向OCR的对抗样本攻击目前仍然存在很多关键的问题需要解决，主要包括以下几个方面：

1）黑盒攻击。尽管对抗样本生成技术在多个领域有潜在的应用，但现存研究在黑盒中的应用相对有限。在分类模型中，黑盒的查询内容相对直接，但对于序列任务，无法进行单步输出，这使得在更少的查询环境中进行更有效的攻击成为一个具有挑战性的问题。

2）攻击方法泛化性弱。以往的攻击研究多基于单一图像，对抗样本攻击的有效性高度依赖于特定输入数据。对于不同的输入图像，需重新构建对抗样本，由于每幅图像的特征独特性，这限制了对抗样本的通用性和在不同数据集或实际场景中的应用范围。这也要求研究人员在设计攻击和防御策略时，需考虑不同的输入数据、模型架构和应用领域。

3）对于汉字的研究相对较少。在字体生成领域，汉字的研究取得了一些成果，但在对抗攻击方面，研究依然偏重于英文攻击。

4.2 未来的发展方向

1）通用攻击方法。以往的研究主要聚焦于基于单一图像的攻击，未来的探索会逐渐转向更复杂的场景，如多语言、多字体或多样式的文本。

2）对抗训练与模型安全性。通过将对抗样本纳入训练集，进一步提高OCR等模型的抵抗攻击能力，确保OCR模型的安全性和可靠性。

3）物理域对抗样本生成。尽管对抗样本技术日益先进，但相应的防御策略研究却稍显滞后。OCR的对抗防御手段仍然不足，而物理域的对抗样本更是为人工智能（artificial intelligence，AI）带来巨大考验，特别是在自动驾驶、安防监控、增强现实和机器人技术等应用领域。这些扰动可能危及系统的安全

和精确性。因此，如何有效识别、抵御对抗攻击，开发更鲁棒的系统变得至关重要。

4）大模型的对抗鲁棒性。GPT-4（generative pre-trained Transformer 4）等大型视觉语言模型（large vision-language models，VLMs）在响应生成方面取得了前所未有的性能，特别是在处理视觉输入的能力上，与专注于文本生成模型相比，能够实现更具创造性和适应性的交互（严昊 等，2023）。被誉为能够分割一切的SAM（segment anything model）图像分割大模型，经过对抗样本的攻击后可能将无法继续保持这种能力。Zhao等人（2023）针对CLIP和BLIP等预训练模型制作有针对性的对抗样本，并迁移到其他VLMs，例如Img2Prompt、UniDiffuser、BLIP-2、MiniGPT-4和LLaVA，表明大模型也受到对抗样本的威胁。如何确保大模型的安全性、隐私性，并充分发挥其潜力，仍是值得进一步探讨的问题。

5 结 语

本文对文本识别的对抗样本攻击和防御方法进行了全面的分析与归类，同时总结了常用数据集和文本识别模型。旨在为后续研究提供一个总结性的参考，帮助研究人员更好地理解现有技术并探索新的方法。作为深度学习中的一个快速发展的研究领域，文本识别对抗样本攻击仍处于不成熟的阶段，存在许多尚待探索的研究方向。随着对文本识别对抗样本攻击与防御研究的深入开展，必然会推动OCR相关技术进一步发展，为OCR的广泛应用提供更安全的保障。

参考文献（References）

Athalye A, Carlini N and Wagner D. 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples//Proceedings of the 35th International Conference on Machine Learning. [s.l.]: PMLR: 274-283

Balkanski E, Chase H, Oshiba K, Rilee A, Singer Y and Wang R. 2020. Adversarial attacks on binary image recognition systems [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/2010.11782.pdf

Baluja S and Fischer I. 2017. Adversarial transformation networks: learning to generate adversarial examples [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1703.09387.pdf

Bartz C, Yang H J and Meinel C. 2017. STN-OCR: a single neural network for text detection and text recognition [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1707.08831.pdf

Bayram S and Barner K. 2023. A black-box attack on optical character recognition systems//Proceedings of 2022 CVMI Computer Vision and Machine Intelligence. Singapore, Singapore: Springer: 221-231 [DOI: 10.1007/978-981-19-7867-8_18]

Borkar T, Heide F and Karam L. 2020. Defending against universal attacks through selective feature regeneration//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 706-716 [DOI: 10.1109/CVPR42600.2020.00079]

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Chen L, Sun J and Xu W. 2020. FAWA: fast adversarial watermark attack on optical character recognition (OCR) systems//Proceedings of 2020 European Conference on Machine Learning and Knowledge Discovery in Databases. Ghent, Belgium: Springer: 547-563 [DOI: 10.1007/978-3-030-67664-3_33]

Chen L and Xu W. 2020. Attacking optical character recognition (OCR) systems with adversarial watermarks [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/2002.03095.pdf

Cipolla R, Gal Y and Kendall A. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7482-7491 [DOI: 10.1109/CVPR.2018.00781]

Deng J C, Dong L, Chen J H, Yan D Q, Wang R D, Ye D P, Zhao L C and Tian J Y. 2023. Universal defensive underpainting patch: making your text invisible to optical character recognition//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: Association for Computing Machinery: 7559-7568 [DOI: 10.1145/3581783.3613768]

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Du X H, Wu H M, Yi Z B, Li S S, Ma J and Yu J. 2021. Adversarial text attack and defense: a review. Journal of Chinese Information Processing, 35(8): 1-15 (杜小虎, 吴宏明, 易子博, 李莎莎, 马俊, 余杰. 2021. 文本对抗样本攻击与防御技术综述. 中文信息学报, 35(8): 1-15) [DOI: 10.3969/j. issn. 1003-0077.2021.08.001]

Feng X J, Yao H X and Zhang S P. 2019. Focal CTC loss for Chinese optical character recognition on unbalanced datasets. Complexity, 2019: #9345861 [DOI: 10.1155/2019/9345861]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1412.6572.pdf

Guo C, Rana M, Cisse M and van der Maaten L. 2018. Countering adversarial images using input transformations [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1711.00117.pdf

He H B, Chen X Y, Wang C Y, Liu J H, Du B, Tao D C and Qiao Y. 2023. Diff-font: diffusion model for robust one-shot font generation [EB/OL]. [2023-09-26]. http://arxiv.org/pdf/2212.05895.pdf

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network. Computer Science, 14(7): 38-39 [DOI: 10.4140/TCP.n.2015.249]

Huang L C, Yang Y, Deng Y F and Yu Y N. 2015. DenseBox: unifying landmark localization with end to end object detection [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1509.04874.pdf

Jia X J, Wei X X, Cao X C and Foroosh H. 2019. ComDefend: an efficient image compression model to defend adversarial examples//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 6077-6085 [DOI: 10.1109/CVPR.2019.00624]

Jiang Y, Lian Z H, Tang Y M and Xiao J G. 2017. DCFont: an end-to-end deep Chinese font generation system//Proceedings of 2017 SIGGRAPH Asia Technical Briefs. Bangkok, Thailand: Association for Computing Machinery: #22 [DOI: 10.1145/3145749.3149440]

Jiang Y, Lian Z H, Tang Y M and Xiao J G. 2019. SCFont: structure-guided Chinese font generation via deep stacked networks//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 33(1): 4015-4022 [DOI: 10.1609/aaai.v33i01.33014015]

Joren H, Gupta O and Raviv D. 2022. Learning document graphs with attention for image manipulation detection//Proceedings of the 3rd International Conference on Pattern Recognition and Artificial Intelligence. Paris, France: Springer: 263-274 [DOI: 10.1007/978-3-031-09037-0_22]

Kherchouche A, Fezza S A and Hamidouche W. 2022. Detect and defense against adversarial examples in deep learning using natural scene statistics and adaptive denoising. Neural Computing and Applications, 34(24): 21567-21582 [DOI: 10.1007/s00521-021-06330-x]

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Kurakin A, Goodfellow I J and Bengio S. 2017. Adversarial machine learning at scale//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR

Lei Q, Wu L F, Chen P Y, Dimakis A G, Dhillon I S and Witbrock M. 2019. Discrete adversarial attacks and submodular optimization with applications to text classification [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1812.00151.pdf

Li M L, Zhong N, Zhang X P, Qian Z X and Li S. 2022. Object-oriented backdoor attack against image captioning//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 2864-2868 [DOI: 10.1109/ICASSP43922.2022.9746440]

Li R, Zheng S Y, Zhang C, Duan C X, Wang L B and Atkinson P M. 2021. ABCNet: attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 181: 84-98 [DOI: 10.1016/j.isprsjprs.2021.09.005]

Li X L, Yu N H, Zhang X P, Zhang W M, Li B, Lu W, Wang W and Liu X L. 2021. Overview of digital media forensics technology. Journal of Image and Graphics, 26(6): 1216-1226 (李晓龙, 俞能海, 张新鹏, 张卫明, 李斌, 卢伟, 王伟, 刘晓龙. 2021. 数字媒体取证技术综述. 中国图象图形学报, 26(6): 1216-1226) [DOI: 10.11834/jig.210081]

Lian Z H, Zhao B and Xiao J G. 2016. Automatic generation of large-scale handwriting fonts via style learning//Proceedings of 2016 SIGGRAPH Asia Technical Briefs. Macau, China: Association for Computing Machinery: #12 [DOI: 10.1145/3005358.3005371]

Liao F Z, Liang M, Dong Y P, Pang T Y, Hu X L and Zhu J. 2018. Defense against adversarial attacks using high-level representation guided denoiser//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1778-1787 [DOI: 10.1109/CVPR.2018.00191]

Liao M H, Shi B G, Bai X, Wang X G and Liu W Y. 2017. TextBoxes: a fast text detector with a single deep neural network//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI: #11196 [DOI: 10.1609/aaai.v31i1.11196]

Liao M H, Zou Z S, Wan Z Y, Yao C and Bai X. 2023. Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 919-931 [DOI: 10.1109/TPAMI.2022.3155612]

Lin J W, Hong C Y, Chang R I, Wang Y C, Lin S Y and Ho J M. 2015. Complete font generation of Chinese characters in personal handwriting style//Proceedings of the 34th IEEEInternational Performance Computing and Communications Conference (IPCCC). Nanjing, China: IEEE: 1-5 [DOI: 10.1109/PCCC.2015.7410321]

Liu X B, Liang D, Yan S, Chen D G, Qiao Y and Yan J J. 2018. FOTS: fast oriented text spotting with a unified network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5676-5685 [DOI: 10.1109/CVPR.2018.00595]

Liu Y H, Cao F M and Zhang Y Q. 2022. Generative adversarial examples for sequential text recognition models with artistic text style//Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods. Science and Technology Publications: 71-79 [DOI: 10.5220/0010866800003122]

Luo C J, Lin Q X, Liu Y L, Jin L W and Shen C H. 2021. Separating content from style using adversarial learning for recognizing text in the wild. International Journal of Computer Vision, 129(4): 960-976 [DOI: 10.1007/s11263-020-01411-1]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2019.

Towards deep learning models resistant to adversarial attacks [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1706.06083.pdf

Moosavi-Dezfooli S M, Fawzi A, Fawzi O and Frossard P. 2017. Universal adversarial perturbations//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 86-94 [DOI: 10.1109/CVPR.2017.17]

Mustafa A, Khan S H, Hayat M, Shen J B and Shao L. 2020. Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing, 29: 1711-1724 [DOI: 10.1109/TIP.2019.2940533]

Odena A, Olah C and Shlens J. 2017. Conditional image synthesis with auxiliary classifier GANs//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 2642-2651 [DOI: 10.5555/3305890.3305954]

Prakash A, Moran N, Garber S, DiLillo A and Storer J. 2018. Deflecting adversarial attacks with pixel deflection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8571-8580 [DOI: 10.1109/CVPR.2018.00894]

Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1804.02767.pdf

Ren H L, Huang T and Yan H Y. 2021. Adversarial examples: attacks and defenses in the physical world. International Journal of Machine Learning and Cybernetics, 12(11): 3325-3336 [DOI: 10.1007/s13042-020-01242-z]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Shi B G, Bai X and Yao C. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(11): 2298-2304 [DOI: 10.1109/TPAMI.2016.2646371]

Shi B G, Wang X G, Lyu P Y, Yao C and Bai X. 2016. Robust scene text recognition with automatic rectification//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4168-4176 [DOI: 10.1109/CVPR.2016.452]

Shi B G, Yang M K, Wang X G, Lyu P Y, Yao C and Bai X. 2019. ASTER: an attentional scene text recognizer with flexible rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2035-2048 [DOI: 10.1109/TPAMI.2018.2848939]

Song C Z and Shmatikov V. 2018. Fooling OCR systems with adversarial text images [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1802.05385.pdf

Su J W, Vargas D V and Sakurai K. 2019. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5): 828-841 [DOI: 10.1109/TEVC.2019.2890858]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R. 2014. Intriguing properties of neural networks [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1312.6199.pdf

Tai J W, Li Y K, Jia X Q and Huang Q J. 2022. A survey: attacks and countermeasures of adversarial examples for speech recognition system. Journal of Cyber Security, 7(5): 51-64 (台建玮, 李亚凯, 贾晓启, 黄庆佳. 2022. 语音识别系统对抗样本攻击及防御综述. 信息安全学报, 7(5): 51-64) [DOI: 10.19363/J.cnki.cn10-1380/tn.2022.09.05]

Telea A. 2004. An image inpainting technique based on the fast marching method. Journal of Graphics Tools, 9(1): 23-34 [DOI: 10.1080/10867651.2004.10487596]

Tesseract. 2016. [EB/OL]. [2023-07-04]. https://github.com/tesseract-ocr/tesseract

Tian Y C. 2016. Rewrite: neural style transfer for Chinese fonts [EB/OL]. [2023-07-04]. https://github.com/kaonashi-tyc/Rewrite

Tian Z, Huang W L, He T, He P and Qiao Y. 2016. Detecting text in natural image with connectionist text proposal network//Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, the Netherlands: Springer: 56-72 [DOI: 10.1007/978-3-319-46484-8_4]

Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D and McDaniel P. 2020. Ensemble adversarial training: attacks and defenses [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1705.07204.pdf

Tramèr F, Papernot N, Goodfellow I, Boneh D and McDaniel P. 2017. The space of transferable adversarial examples [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1704.03453.pdf

Wan Z Y, Xie F M, Liu Y B, Bai X and Yao C. 2019. 2D-CTC for scene text recognition [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1907.09705.pdf

Wang W H, Xie E Z, Li X, Liu X B, Liang D, Yang Z B, Lu T and Shen C H. 2022. PAN++: towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 5349-5367 [DOI: 10.1109/TPAMI.2021.3077555]

Wick C, Reul C and Puppe F. 2018. Calamari—A high-performance tensorflow-based deep learning package for optical character recognition [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/1807.02004.pdf

Wu S J, Yang C Y and Hsu J Y J. 2020. CalliGAN: style and structure-aware Chinese calligraphy character generator [EB/OL]. [2023-07-04]. http://arxiv.org/pdf/2005.12500.pdf

Xiao C W, Li B, Zhu J Y, He W, Liu M Y and Song D. 2018. Generating adversarial examples with adversarial networks//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 3905-3911 [DOI: 10.24963/ijcai.2018/543]

Xie C H, Wu Y X, van der Maaten L, Yuille A L and He K M. 2019. Feature denoising for improving adversarial robustness//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 501-509 [DOI: 10.1109/CVPR.2019.00059]

Xie Y C, Chen X Y, Sun L and Lu Y. 2021. DG-Font: deformable generative networks for unsupervised font generation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5130-5140 [DOI: 10.1109/CVPR46437.2021.00509]

Xu S H, Jin T, Jiang H and Lau F C M. 2009. Automatic generation of personal Chinese handwriting by capturing the characteristics of personal handwriting//Proceedings of the 21st Innovative Applications of Artificial Intelligence Conference. Pasadena, USA: AAAI: 191-196

Xu X, Chen J F, Xiao J H, Gao L L, Shen F M and Shen H T. 2020a. What machines see is not what they get: fooling scene text recognition models with adversarial text images//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 12301-12311 [DOI: 10.1109/CVPR42600.2020.01232]

Xu X, Chen J F, Xiao J H, Wang Z, Yang Y and Shen H T. 2020b. Learning optimization-based adversarial perturbations for attacking sequential recognition models//Proceedings of the 28th ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery: 2802-2822 [DOI: 10.1145/3394171.3413543]

Xu Y K, Dai P W and Cao X C. 2021. Less is better: fooling scene text recognition with minimal perturbations//Proceedings of the 28th International Conference on Neural Information Processing. Sanur, Indonesia: Springer: 537-544 [DOI: 10.1007/978-3-030-92310-5_62]

Xu Y K, Dai P W, Li Z K, Wang H J and Cao X C. 2023. The best protection is attack: fooling scene text recognition with minimal pixels. IEEE Transactions on Information Forensics and Security, 18: 1580-1595 [DOI: 10.1109/TIFS.2023.3245984]

Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(9): 2749-2762 (严昊, 刘禹良, 金连文, 白翔. 2023. 类 ChatGPT 大模型发展、应用和前景. 中国图象图形学报, 28(9): 2749-2762) [DOI: 10.11834/jig.230536]

Yang M K, Zheng H T, Bai X and Luo J B. 2021. Cost-effective adversarial attacks against scene text recognition//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE: 2368-2374 [DOI: 10.1109/ICPR48806.2021.9412914]

Yang S, Wang Z Y, Wang Z W, Xu N, Liu J Y and Guo Z M. 2019. Controllable artistic text style transfer via shape-matching GAN//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4441-4450 [DOI: 10.1109/ICCV.2019.00454]

Yuan L, Li X M, Pan Z X, Sun J M and Xiao L. 2022. Review of adversarial examples for object detection. Journal of Image and Graphics, 27(10): 2873-2896 (袁珑, 李秀梅, 潘振雄, 孙军梅, 肖蕾. 2022. 面向目标检测的对抗样本综述. 中国图象图形学报, 27(10): 2873-2896) [DOI: 10.11834/jig.210209]

Yuan S Z, Liu R X, Chen M, Chen B Y, Qiu Z J and He X D. 2022. SE-GAN: skeleton enhanced gan-based model for brush handwriting font generation//Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei, China: IEEE: 1-6 [DOI: 10.1109/ICME52920.2022.9859964]

Yuan X Y, He P, Lit X and Wu D P. 2020. Adaptive adversarial attack on scene text recognition//Proceedings of 2020 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Toronto, Canada: IEEE: 358-363 [DOI: 10.1109/INFOCOMWKSHPS50562.2020.9162685]

Zhang T, Yang K W, Wei J H, Liu Y and Ning Y L. 2022. Survey on detecting and defending adversarial examples for image data. Journal of Computer Research and Development, 59(6): 1315-1328 (张田, 杨奎武, 魏江宏, 刘扬, 宁原隆. 2022. 面向图像数据的对抗样本检测与防御技术综述. 计算机研究与发展, 59(6): 1315-1328) [DOI: 10.7544/issn1000-1239.20200777]

Zhong D J, Lyu S J, Shivakumara P, Yin B, Wu J J, Pal U and Lu Y. 2022a. SGBANet: semantic GAN and balanced attention network for arbitrarily oriented scene text recognition//Proceedings of the 17th European Conference on Computer Vision (ECCV 2022). Tel Aviv, Israel: Springer: 464-480 [DOI: 10.1007/978-3-031-19815-1_27]

Zhong Y H, Cheng X Y, Chen T, Zhang J, Zhou Z K and Huang G. 2022b. PRPN: progressive region prediction network for natural scene text detection. Knowledge-Based Systems, 236: #107767 [DOI: 10.1016/j.knosys.2021.107767]

Zhou B Y, Wang W H and Chen Z H. 2011. Easy generation of personal Chinese handwritten fonts//Proceedings of 2011 IEEE International Conference on Multimedia and Expo. Barcelona, Spain: IEEE: 1-6 [DOI: 10.1109/ICME.2011.6011892]

Zhu H G, Zhu Y, Zheng H R, Ren Y C and Jiang W M. 2023. LIGAA: Generative adversarial attack method based on low-frequency information. Computers and Security, 125: #103057 [DOI: 10.1016/j.cose.2022.103057]

作者简介

郭凯威，男，硕士研究生，主要研究方向为大数据与智能安全。E-mail: sguokaiwei@163.com

杨奎武，通信作者，男，副教授，主要研究方向为大数据与智能安全。E-mail: yangkw@aliyun.com

张万里，男，博士研究生，主要研究方向为深度学习和大数据安全。E-mail: wanli_zhang@aliyun.com

胡学先，男，副教授，主要研究方向为应用密码学和大数据安全。E-mail: xuexian_hu@hotmail.com

刘文钊，男，博士研究生，主要研究方向为大数据安全。E-mail: liuwz_2000@hotmail.com