IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2024, accepted 4 August 2024, date of publication 13 August 2024, date of current version 22 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3442976

SURVEY

Image CAPTCHAs: When Deep Learning Breaks the Mold

MORTEZA MORADI[1], MOHAMMAD MORADI[2], SIMONE PALAZZO[1], FRANCESCO RUNDO[3], AND CONCETTO SPAMPINATO[1]

[1]PeRCeiVe Lab, University of Catania, 95124 Catania, Italy
[2]Department of Electrical, Electronics and Computer Engineering, University of Catania, 95124 Catania, Italy
[3]STMicrolectronics, ADG Central Research and Development, 95121 Catania, Italy

Corresponding author: Morteza Moradi (morteza.moradi@phd.unict.it)

ABSTRACT While text-based CAPTCHAs have been the predominant type of human interaction proofs (HIPs) for many years, image recognition challenges have also gained significant attention. This trend is due, on one hand, to groundbreaking advancements in solving text CAPTCHAs and, on the other hand, to the intrinsic weakness of machines in dealing with cognitive tasks such as those presented in image CAPTCHAs. In addition to classic and even human-centric image CAPTCHA solvers, deep learning has recently emerged as a significant player, providing two unprecedented and, at the same time, contradictory advantages for designers and adversaries. Designers benefit from deep learning techniques to make CAPTCHAs as hard to break as possible, while adversaries utilize deep learning algorithms to attack novel and complicated image-based challenges. Given these premises, this paper presents an analytical study on the applications of deep learning for and against image CAPTCHAs. This study aims to provide a comprehensive overview of the latest advancements in the field, assisting researchers and practitioners in designing image CAPTCHAs that are both user-friendly and resilient against modern attacks.

INDEX TERMS CAPTCHA protection, deep learning, image CAPTCHA, object recognition.

I. INTRODUCTION

Due to numerous usability and accessibility issues, the Completely Automated Public Turing test to tell Computers and Humans Apart, or simply CAPTCHA [1], is not a popular security mechanism among users. However, as there is still no reliable and widely accepted alternative, the vast majority of high-traffic websites [2] currently use CAPTCHA to distinguish between humans and bots attempting to imitate human behavior. Among different types of CAPTCHA, image-based CAPTCHAs have gained much attention in recent years [2]. Two main reasons contribute to this popularity are the vulnerability of common text CAPTCHAs to deep learning-based attacks [3], [4], and the intrinsic weakness of machines in dealing with cognitive challenges [5], [6], [7]. Attractive aesthetics and enhanced user experience are also factors influencing the recent preference for image CAPTCHAs. Although a wide variety of image-based CAPTCHAs have been proposed over the years, only some have gained enough popularity to become a serious choice for real-world applications. However, these methods are exposed to many attacks by adversaries aiming to break image-based challenges for malicious purposes and vulnerability identification. Early works in this area mainly relied on standard computer vision and shallow machine learning techniques. However, with the emergence of advanced image CAPTCHAs, since the introduction of reCAPTCHA v2 in 2014, more sophisticated breaking methods based on deep learning have been designed. On the other hand, designers also leverage deep learning capabilities to increase the robustness of CAPTCHAs, for example, by introducing adversarial CAPTCHAs. Given the importance of CAPTCHA both as a security measure and as a benchmark task to measure artificial intelligence's human-likeness, this paper aims to provide an analytical study on the applications and impacts of deep learning-based methods

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

for breaking and securing image CAPTCHAs. In addition to investigating the latest advancements in the field, early works are also considered to better reflect two decades of developments in CAPTCHA research. The ideas and concepts presented in this paper are intended to provide researchers and practitioners in cybersecurity and computer vision with insights into current threats and attacks on CAPTCHA systems. At the same time, the introduced techniques can be utilized to strengthen CAPTCHA tests and enhance their robustness against sophisticated attacks. The remainder of the paper is organized as follows: To underscore the significance of text-based Human Interactive Proofs (HIPs), Section II provides a succinct review of recent advancements in text CAPTCHA breaking techniques. The motivation for this research study is introduced in section III. Different methods of breaking image CAPTCHAs, including classical and deep learning-based approaches, are thoroughly studied in section IV and its subsections. To conduct the main part of the review, all published papers with original contributions on using deep learning for breaking image CAPTCHAs are selected. Moreover, to provide an overview of classical machine learning and computer vision based techniques for breaking image CAPTCHAs, we discuss some of the most cited works with novel underlying ideas as a source of inspiration for future works and as a showcase of advancements in the field over the years. Investigation into how deep learning can be used to protect image CAPTCHAs against modern attacks is conducted in section V. Finally, in section VI, we provide discussions on different aspects of image CAPTCHA breaking and suggestions for future research in the field.

II. STATE-OF-THE-ART IN CAPTCHA BREAKING

CAPTCHA, as a controversial phenomenon, has garnered significant attention from various parties, including general Internet users, researchers, and gray/black hat adversaries. While adversarial parties may attempt to crack CAPTCHAs for amusement, financial gain, or other malicious purposes, researchers view these reverse Turing tests as challenging tasks to assess the capabilities and reliability of artificial intelligence in solving difficult problems [8]. Moreover, these endeavors have proven useful for identifying vulnerabilities in CAPTCHA tests and presenting new ideas to enhance their robustness against adversarial attacks. Therefore, research on different aspects of CAPTCHAs, specifically proposing breaking techniques, has become a focal point over the years. In this section, we present a brief overview of recent advances in breaking text CAPTCHAs with a twofold objective. On one hand, it offers a quick snapshot of the state-of-the-art deep learning-based methods, which may inspire the design of attacks on image CAPTCHAs. On the other hand, it is common for modern image CAPTCHAs to present users with distorted text descriptions, technically similar to the challenges posed by text-based CAPTCHAs. Consequently, the introduced methods can be considered as part of the process of breaking image CAPTCHAs. As text CAPTCHAs dominated the landscape for years, most initial efforts were devoted to breaking these CAPTCHAs. In addition to early works [9], [10], recent attempts leveraging deep learning-based techniques have yielded unprecedented precision and accuracy.

One of the most important models in the field that characterized by exploiting concepts from cognitive neuroscience, such as cortical function, is the Recursive Cortical Network (RCN) [11]. It consists of a hierarchical model where objects are represented through a combination of contours and surfaces. Besides breaking text CAPTCHAs, it has been employed for various visual recognition tasks, including scene text recognition and occlusion reasoning. In terms of performance for targeting reCAPTCHA v1, RCN achieved 94.3% accuracy for character recognition and 66.6% accuracy for word recognition. To solve complex text CAPTCHAs, the authors in [12] take advantage of a GAN-based transformation module to convert complicated CAPTCHAs into simpler instances. This strategy ideally facilitates the tasks of character segmentation and recognition. The proposed recognition network follows a straightforward CNN structure, including two convolutional blocks, each consisting of two convolutional layers. The obtained results of 96% accuracy for character recognition and 74% accuracy for solving the CAPTCHA test demonstrate the effectiveness of the proposed model. In [13], the performance of three CAPTCHA-solver architectures, namely CNN with Bidirectional LSTM, CNN with Decoder Transformer, and patch-based single Transformer, has been examined. In the first architecture, a CNN with four convolution layers along with batch normalization serves as an encoder, and its output is passed, as a sequential input, to four stacked layers of a Bidirectional LSTM. The second architecture is similar to the first one, but it reduces the number of training parameters, employing a single Transformer network as a decoder, while the third model consists of only a single Transformer network. The experimental results show that the Transformer-based model yields the best accuracy over all the five datasets used in the study. In contrast to the works that use manually annotated training data, [14] presents an approach based on training CAPTCHA solvers with an automatically created dataset. The underlying idea is to combine classic brute-force attack with transfer learning. For the character segmentation part, a basic method (using Otsu binarization, morphological noise reduction, and finally separating overlapped characters) is proposed. On the other hand, for the recognition task, a small training dataset containing 500 syntactic single-character images is used. The best performance of the proposed method is achieved when targeting a 5-digit CAPTCHA with 94.78% accuracy. However, when it comes to CAPTCHAs of the same length made up of lowercase alphabetical characters, the accuracy drops to 32.89%. Attempting to break four-character CAPTCHAs, an end-to-end CNN-RNN model without the need for any preprocessing and character segmentation step is proposed in [15]. In this model, a CNN

network extracts features from input character images, and then a classification step is carried out using an RNN network. This ResNet-GRU model achieves an accuracy of over 99% on three of the four tested datasets. Faster R-CNN [16] for breaking text-based CAPTCHAs was investigated in [17], but despite its effectiveness, it fails with small-size characters. In an attempt to address this problem, the authors in [18], in addition to adopting Fast R-CNN, proposed a new Feature Refine Network (FRN) to take advantage of multi-scale feature fusion to learn features at different scales and resolutions. Therefore, skip ROI pooling has been used to facilitate feature extraction at multiple scales. This model could break Hotmail, Baidu, and eBay CAPTCHAs with 94.2%, 96.8%, and 97.3% accuracy, respectively.

The main idea presented in [19] lies in separating the font structure and font style of characters in handwritten CAPTCHAs. Indeed, the authors further observed that the character recognition process significantly depends on the font structure. The proposed framework consists of two major parts: 1) a style transfer network, partially inspired by U-Net [20], DCGAN [21], and work done in [22], for converting stylized characters into their standard forms, and 2) a recognition network to identify each character and consequently break the CAPTCHA. The reported performance shows an accuracy of 89% for breaking reCAPTCHA v1 and 88% for eBay CAPTCHA. As another contribution in the field, [23] leverages Capsule Networks that take into account spatial relationships among object parts' features for breaking digit CAPTCHAs. Despite its high computational burden, it yields performance comparable to (and in some scenarios better than) CNN-based CAPTCHA solvers.

Nevertheless, most of the investigated methods are originally designed for CAPTCHAs with English characters. Intrinsic differences between various languages, specifically in size, variation, and representation of alphanumeric characters (e.g., as in Persian/Arabic and Chinese languages), make it hardly possible (if not totally impractical) to design a language-agnostic text CAPTCHA solver. In this regard, some studies on breaking non-English text CAPTCHAs have been carried out [24], [25], [26]. For example, an interesting work is proposed in [27], where a deep Siamese network, using residual connections, for one-shot and few-shot Chinese CAPTCHA recognition is presented. The problem in this work is formulated as a binary classification task, where a pair of character images are received as input, and the model decides if they are the same characters.

III. RESEARCH OBJECTIVES AND MOTIVATIONS

Image CAPTCHAs, generally, are tests that ask users to perform an image recognition task to prove their genuineness. Due to the wide variety of proposed image CAPTCHAs, introducing a comprehensive classification may not be trivial. Nonetheless, a tentative categorization foresees five types of methods: face recognition [28], [29], image annotation [30], [31], image selection [32], image matching [33], and puzzle-based ones [34]. Most image-based CAPTCHAs share some level of similarity (since users should rely on their visual cognition ability to interpret and solve the challenge), but differences in interaction modality and presentation may vary. However, only a few instances have gained enough momentum to be widely employed in real-world applications. This is, of course, due to both design and implementation considerations and feasibility concerns. In this regard, despite introducing some interesting ideas for image-based CAPTCHAs, most CAPTCHA proposals remain stuck at the conceptual stage and have hardly been used in any real-life applications. Those that have found real applications share similar features, including workflow, modalities, and even appearance. Such commonality and best practices that can be seen in the design and development of principal CAPTCHAs (as in Figure 1) help users to interact with different tests without facing remarkable irregularity.

The three famous CAPTCHAs illustrated in Figure 1 are good examples of modern image CAPTCHAs. The tests provide text instructions, and users should select images matching the description. In contrast with reCAPTCHA and hCaptcha, GeeTest uses graphical elements rather than real-world images, and it can also improve its security.

Indeed, breaking CAPTCHAs, aside from potential commercial or unethical motivations, serves as a means to gauge the ability of intelligent machines to emulate human-level cognitive intelligence and behavior. Taking a closer look at the literature reveals that, thanks to deep learning, solving text- and digit-based CAPTCHAs is now considered a doable task. Nevertheless, when it comes to more challenging cognitive-intensive tasks, such as image recognition and understanding, there is still room for improvement. This study tackles a notable gap in current literature, lacking a comprehensive exploration to serve as a benchmark for incorporating the latest advances in deep learning aimed at compromising CAPTCHA security, while also bolstering resistance against breaking attempts. Indeed, much of the research in the field has been directed towards devising new CAPTCHA challenges or utilizing deep learning to bypass existing schemes. However, the potential of deep learning-based methods for safeguarding CAPTCHAs, particularly image CAPTCHAs, is somewhat undervalued. Hence, the aim is to emphasize this significant gap and promising research avenue, shedding light on strategies for reinforcing image CAPTCHAs against contemporary, potent attacks. This is especially pertinent given that, despite the emergence of numerous effective security measures for web-based systems, CAPTCHAs remain pivotal in a wide array of practical applications. Accordingly, the main objective of this paper is to provide an analytical overview of the state of the art of deep learning-based methods for both breaking image CAPTCHAs. Also, aiming to report the progress in the field and provide some inspiration for future works, classic attacks on image CAPTCHAs (mainly those based on machine learning) are also investigated. Moreover, benefits

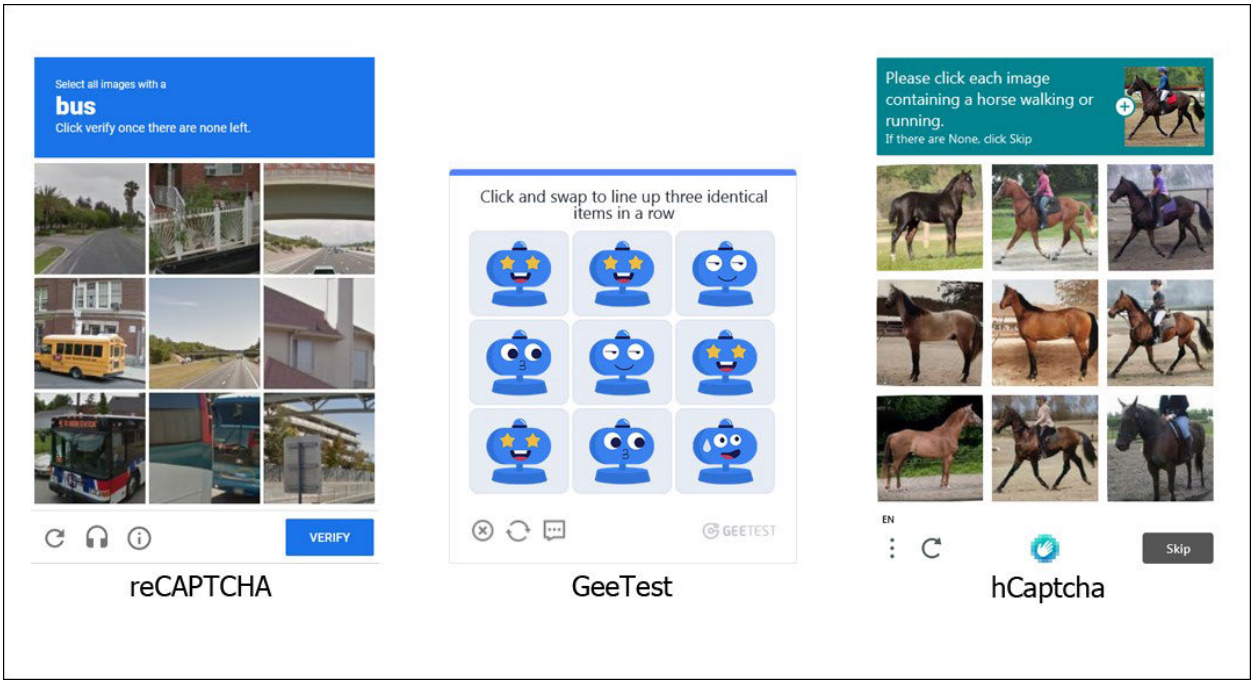


FIGURE 1. Similar appearance of three modern image-based CAPTCHAs.

and opportunities deep learning puts forward for securing and protecting image CAPTCHAs are discussed.

It is worth mentioning that, despite the publication of several review papers on CAPTCHA breaking methods, our contribution stands out distinctly from them. Specifically, the work by Zhang et al. [35] provided a brief overview of various methods for designing and attacking CPATCHAs but did not focus only on image CAPTCHAs. Similarly, survey papers by Xu et al. [36] and Guerar et al. [37] provided general studies on the field's progress, including notes on attacks on different types of CAPTCHA. Additionally, Kumar et al. [38] conducted another survey and, while providing more information on different attacks on CAPTCHAs, including image-based ones, did not include details on state-of-the-art attacks targeting modern image CAPTCHAs. The recent survey paper by Tariq et al. [39] also discussed CAPTCHA breaking techniques concisely, without delving into an in-depth analysis of those methods. With this understanding, our research distinguishes itself as the inaugural study to explore deep learning-based attacks and securing mechanisms for image CAPTCHAs.

Additionally, this research is subject to several limitations. Foremost among these is the restricted availability of data and program codes pertaining to certain attack scenarios and image CAPTCHAs. This limitation impedes the ability to provide comprehensive comparisons across various facets of attacks, thereby partially limiting the precision of performance analyses. Moreover, the absence of universally accepted evaluation criteria or benchmarks further constrains the scope of comparative assessments. Nevertheless, these challenges stem from the diverse nature of (image) CAPTCHAs and the corresponding array of attacks. Considering the aforementioned constraints, the primary objective of this study is to investigate the deep learning-based attack methodologies on image CAPTCHAs and assess their respective success rates. Rather than attempting direct comparisons of performance superiority among diverse methods, which may prove impractical given the breadth of this study.

IV. BREAKING IMAGE CAPTCHAS

The significance of image CAPTCHA recognition lies in its distinctiveness from other types of CAPTCHAs. Generally, breaking text-based CAPTCHAs follows a series of steps, including pre-processing and noise reduction, feature extraction, character segmentation, classification, and possibly a final refinement step, despite minor variations based on factors such as language-specific requirements and protection mechanisms [4]. Similarly, the process for audio-based CAPTCHAs shares common steps, involving transcribing a noisy audio track to text and solving the test [40], [41]. However, image CAPTCHAs present a different challenge due to the variety of tests and solving approaches they employ. This diversity makes it impractical to design a single, general solver for all image CAPTCHAs. Consequently, exploring the methodologies for breaking various types of image CAPTCHAs becomes an intriguing topic. Understanding how different image CAPTCHAs can be broken sheds light on the development of artificial intelligence and computer vision techniques. Moreover, it reveals vulnerabilities in current CAPTCHAs, offering insights for enhancing their reliability and resilience against machine-driven attacks. Today's image CAPTCHAs are more sophisticated than their predecessors, evolving in response to advancements

in CAPTCHA-breaking techniques. Traditional adversarial strategies against CAPTCHAs have proven ineffective and impractical due to employing latest protective measures, which have even thwarted human-in-the-loop attacks [42]. Therefore, to delve into recent advancements, the subsequent section reviews deep learning-based attacks and breaking methods. The analysis begins by examining classic attack strategies.

A. TRADITIONAL ATTACKS

Although it is challenging to precisely chronicle the early developments in (image) CAPTCHAs, there is no doubt that ARTiFACIAL [43] stands out as one of the pioneering image CAPTCHAs. Drawing inspiration from text-based CAPTCHAs, ARTiFACIAL creates an image featuring a distorted face set against a cluttered and highly noisy background. To validate its authenticity, users are required to locate the face and identify six key facial points (four eye corners and two mouth corners). This test marked a significant breakthrough, aligning with the core principle of CAPTCHAs: ease of solving for humans and extreme complexity for machines. As reported in the original paper, human participants achieved a remarkable 99.7 To assess the robustness of this CAPTCHA, it underwent testing against three state-of-the-art face detection methods. The first detector [44] employed information-based maximum discrimination (MD). The second utilized a sparse network of linear functions, known as SNoW [45], while the third face detector [46] relied on Adaboost with a cascade of linear features. These three attacks were conducted on 1000 challenges (synthesized images), resulting in an exceptionally low performance, with accuracy levels approaching zero percent.

A decade later, another computer vision-based attack was proposed to break the ARTiFACIAL CAPTCHA [47]. In contrast to the approach in [43], which relied on general face detection algorithms, the method outlined in [47] introduces a case-specific strategy to contend with the noise and misleading shapes embedded within the CAPTCHA. This attack comprises two steps: face detection and facial feature extraction. In the initial stage, a gradient face detector is trained on sets of gradient faces and non-faces using a boosting chain, followed by the removal of vertical and horizontal lines from the gradient image. In the subsequent phase, an intensity correction is first applied to the detected face to facilitate the identification of facial components. Subsequently, a face alignment algorithm employing component-based discriminative search [48] is utilized to pinpoint the precise positions of crucial facial elements required to break the challenge. The effectiveness of this approach was assessed over 800 images, resulting in an overall success rate of 18% in breaking the ARTiFACIAL CAPTCHA, a significant improvement over the methods outlined in [43]. Furthermore, the response time in this study dramatically decreased to 1.47 seconds from the 14 seconds reported in [43].

FIGURE 2. An example of Asirra CAPTCHA [32].

Asirra [32] can be viewed as the precursor to modern image selection CAPTCHAs. It tasks users with identifying cat images among a set of 12 pictures containing both cats and dogs (refer to Figure 2). As outlined in the original paper, the test was easily solved by humans, achieving a success rate of 99.6%. To gauge the robustness of Asirra, the authors posited that if attackers utilized classifiers with 60% accuracy, they could break a 12-image test with approximately 21% accuracy.

To circumvent Asirra, researchers proposed an SVM classifier trained on color and texture features of cat and dog images in [49]. However, the reported success rate on a dataset of approximately 13,000 images was only 10.3%, demonstrating inferior performance compared to previous attempts [49]. Another approach, detailed in [50], utilized a combination of features extracted with the ENT tool [51], such as size, entropy, and mean, alongside a LogitBoost classifier [52], achieving a success rate slightly above 58%. Additionally, an attack leveraging the Hierarchical Temporal Memory (HTM) network proposed in [53] attained an accuracy of 74.7% using a dataset comprising 12,500 images.

Another notable image CAPTCHA, IMAGINATION [54] (refer to Figure 3), presents users with a synthesized image consisting of eight sub-images. Users are tasked with clicking around the geometric center of one of these sub-images and assigning a corresponding label. If the clicked area is close to one of the centers, an additional distortion is applied to the corresponding sub-image, and the user is prompted to select from provided label options to describe the image. The authors reported a human success rate of 95%. However, a computer vision-based attack outlined in [55]achieved a breaking accuracy of 74.31%. The attack's initial step involves detecting rectangular regions within the composite CAPTCHA image to delineate boundaries of each of the eight sub-images. Subsequently, a Gaussian filter is applied for noise reduction, followed by color edge and line segment detection. Candidate rectangles are then ranked

FIGURE 3. The composite image (challenge) of IMAGINATION CAPTCHA [54].

based on various criteria, including edge intensity. Finally, the consistency of identified rectangles is assessed using a priori knowledge to identify overlapping sub-images and select one for solving the challenge.

In [56], novel attack strategies are presented targeting two face recognition CAPTCHAs: FaceDCAPTCHA [57] and FR-CAPTCHA [58]. To break the former, the method initially employs edge detection to segment face and non-face images, followed by extraction of color and texture features, including SIFT, LBP, and Laws' Masks. These features are then utilized to train an SVM classifier to predict real face images. For the FR-CAPTCHA, where the challenge involves identifying two face images of the same person, a Haar classifier is first used to detect real human faces, followed by a feature extraction process. Classification is then performed by calculating the Euclidean distance between feature vectors to determine the pair with the smallest distance as the answer. The proposed methods achieved accuracies of 48% and 23% in breaking FaceDCAPTCHA and FR-CAPTCHA, respectively.

In a study outlined in [59], an attack named SliAttack on slide-based CAPTCHAs, also known as puzzle CAPTCHAs, is introduced. In these CAPTCHAs, users must place a puzzle piece in the correct position using a mouse. The attack comprises two steps: 1) detection of the puzzle region, and 2) simulating human behavior in completing the task. The first step involves comparing the CAPTCHA challenge with its original image to determine the puzzle's position. The original image can be obtained by analyzing various CAPTCHAs designed based on it, necessitating the collection of a significant number of the target CAPTCHA's challenges. To mimic human behavior, a sigmoid-based simulation function is employed. The attack achieved success rates of 96% on GeeTest, 100% on Tencent, and 98% on Netease CAPTCHAs.

Design flaws and issues with the renowned puzzle-based CAPTCHA, Capy,[1] and the gender classification CAPTCHA, FunCAPTCHA, were investigated by Hernandez-Castro et al. in [60] and [61]. CAPTCHaStar [62], a shape discovery CAPTCHA that requires users to identify shapes within a cluttered environment, was initially breached using an SVM classifier with 78% accuracy within 421 seconds. According to the authors, humans can typically solve the challenge in an average of 27 seconds with a success rate of 90%. Subsequent attacks on CAPTCHaStar, however, managed to break the CAPTCHA with 85% accuracy [63] using the BASECASS methodology [64], and with a 96% success rate through the development of an ad-hoc technique [65], both of which demonstrated remarkable performance.

[1] https://www.capy.me

B. DEEP LEARNING-BASED ATTACKS

In the past decade, remarkable advancements in computer vision and machine learning techniques, along with a thriving underground market for human-based solvers, have exposed the diminishing effectiveness and resilience of traditional CAPTCHA challenges against various attacks. Consequently, numerous efforts have been undertaken to enhance the security of CAPTCHAs against human-based interventions, as evidenced by works such as [66] and [67]. Additionally, several risk analysis engines [68], pioneered by industry leaders like reCAPTCHA, hCaptcha, and Capy, have been introduced to provide intelligent and user-friendly protection. Furthermore, CAPTCHA challenges have evolved towards more cognitive-oriented and consequently intricate tasks for machines. This evolution prompted the adoption of deep learning techniques to combat the heightened security measures integrated into authentication tests. To the best of our knowledge, the inaugural deep learning-based attack on CAPTCHA was unveiled during the ICMLA 2012 Face Recognition Challenge. This initiative targeted Avatar CAPTCHA [69], designed to differentiate between human faces and artificially generated ones. According to reports [70], the attack, leveraging a six-layered convolutional neural network, achieved a staggering 99% accuracy in bypassing the CAPTCHA.

1) ATTACKS ON RECAPTCHA

Google's reCAPTCHA stands out as the most widely utilized image CAPTCHA across the web [2], drawing significant attention from both researchers and potential hackers eager to showcase the capabilities of their deep learning models in circumventing its security. Thus, this section delves into research endeavors directed at breaching reCAPTCHA.

The initial attempt to break reCAPTCHA using deep learning techniques was undertaken by researchers affiliated with the Google Street View and reCAPTCHA teams [71]. Their study aimed to accurately recognize digit numbers from Street View images, approaching the task with the goal of achieving human-level performance. Addressing the challenge of transcribing street numbers, as presented in their version of reCAPTCHA, the authors framed the task as a

specialized form of sequence recognition. Consequently, their objective was to learn $P(S|X)$, where S represents the output sequence and X denotes the input image. To maintain the requisite accuracy, the authors employed a data augmentation process to expand the dataset's size. Achieving a 98% accuracy rate on the Street View House Numbers (SVHN) dataset, closely mirroring human performance, relied on the utilization of a confidence thresholding mechanism. The proposed methodology was also tested on two additional datasets beyond reCAPTCHA. The first dataset, an extended version of SVHN, yielded a 91% accuracy rate across the overall sequence. The second supplementary dataset comprised challenging text-based reCAPTCHA examples, where the proposed model attained an impressive accuracy of 99.8%, demonstrating its generalization capabilities. An essential aspect of this work lies in its implementation, notably the utilization of a single neural network capable of simultaneously performing various tasks, including localization, segmentation, and recognition.

The first deep learning-based effort to bypass Google's No CAPTCHA reCAPTCHA was pioneered by Sivakorn et al. in 2016 [72]. Their proposed CAPTCHA-breaking methodology involves several sequential steps: 1) Extraction of candidate images for each challenge; 2) Utilization of reverse image search to gather supplementary descriptive information and higher-quality versions of images, if available; 3) Leveraging various deep learning-based image annotation systems and frameworks, including Clarifai [73], Alchemy, TDL [74], NeuralTalk [75], and Caffe [76], to obtain reliable tags. Subsequently, a Word2Vec classifier is employed to gauge the similarity between the acquired tags and the informative descriptions provided within the challenge. The effectiveness of the attack, initially employing reverse image search alone, stood at 13.1%. However, by incorporating web page titles associated with the searched images, this success rate increased to 19.2%. Furthermore, when supplemented with additional metadata, such as those provided by deep learning image annotation tools, the results significantly improved. Specifically, utilizing Alchemy yielded a success rate of 49.9%, Clarifai achieved 58%, and Caffe reached 45.9%.

reCAPTCHA v2, also known as the No CAPTCHA reCAPTCHA, offers a variety of challenges, ranging from identifying similar images to selecting squares containing fragments of larger images or concepts. These challenges encompass diverse content categories, such as road signs, traffic scenes, and natural landscapes. Due to this diversity, it is challenging to devise a single solver capable of effectively tackling all variations of these challenges.

Zhou et al. [77] directed their efforts towards overcoming the road sign detection challenge by introducing two distinct Deep Neural Network (DNN) models: 1) The first model utilized GoogleNet [78], pre-trained on ImageNet, with the final three layers replaced by a fully connected layer and an output layer containing only two neurons. 2) The second model was a custom-designed 24-layer convolutional neural network, comprising five sets of convolutions, Rectified Linear Unit (ReLU) activations, normalization and pooling layers, along with a softmax layer. Both models underwent training and validation using a dataset comprising 6,400 road sign images sourced from Flickr and Google, categorized into 16 classes. Results indicated that the CAPTCHA was successfully breached with a 94.25% accuracy rate by the GoogleNet-based model, whereas the customized CNN network achieved an accuracy of 77.2%.

SelAttack [59] was developed to overcome image selection CAPTCHAs, including reCAPTCHA v2 (specifically reCAPTCHA 2015 and 2018). The attack methodology involves the following steps: 1) utilizing an image classifier and a character recognition model to interpret challenge descriptions, 2) extracting candidate images along with their descriptions, and 3) employing a classification model to assign semantic labels to challenge images. Ultimately, images whose assigned labels match the provided descriptions are selected as the answer(s). Two CNN models were specifically tailored for breaking reCAPTCHA 2015 and 2018, trained on datasets comprising 33,000 and 15,000 images, respectively. Reported results indicate an 88% accuracy in solving reCAPTCHA 2015 and a 79% success rate in attacking reCAPTCHA 2018.

Inspired by the work of Sivakorn et al. [72], Alqahtani and Alsulaiman [79] pursued their efforts in breaking reCAPTCHA by exploring various machine learning techniques leveraging a deep learning-based image tagging service called Imagga.[2] In their proposed approach, the sample image (i.e., the image accompanied by the text description) is initially tagged, and the top five tags along with their confidence scores are utilized to describe the sample image. This process is iterated for all candidate images (i.e., each of the nine challenge images). Euclidean distances between the confidence values of the sample's vector and those of the candidate images are calculated to assess semantic similarity. These similarity vectors are then fed into several machine learning classifiers. Results show a 26.57% accuracy for Naive Bayes, 37.57% accuracy using CART, 44.43% for Bagging, and 45.42% for a Random Forest classifier. Furthermore, incorporating context information in the form of the challenge's description led to improved success rates of 56.29% for Random Forest, 54.86% for CART, and 38.43% for the Naive Bayes classifier.

The hybrid breaking technique proposed in [80] combines simulated human mouse movement with deep learning-based object detection. To address the risk analysis engine of the challenge, the strategy involves moving the mouse along a Bezier curve defined by randomly selected points. This tactic is crucial as receiving low scores from the engine, which is likely in the absence of mouse-like activity, can complicate the challenge. Thus, this step serves as an implicit preprocessing phase, making the image-based challenges easier and less noisy for the object detection algorithm. For

[2]https://imagga.com/solutions/auto-tagging

image recognition, the process resembles previous efforts in the field, involving the interpretation of text descriptions and additional information such as HTML code to identify if candidate images have been replaced. Object detection relies on a transfer learning-based approach utilizing two customized models, both based on YOLOv3 [81], and trained on distinct datasets. The first model is trained on a dataset sourced from ImageNet [82] and augmented with results from Google image search for categories absent in ImageNet. Conversely, the second model is trained on real-world reCAPTCHA v2 challenges. The ImageNet-trained model achieved 61% accuracy (38% without mouse movement simulation), while the other model attained 68% accuracy (41% without mouse movement simulation).

The reCAPTCHA solver detailed in [83], comprises two distinct modules: browser automation and solver. The former module focuses on automating browser-based inspection tasks, such as identifying different elements within frames (e.g., checkboxes and challenge submission buttons), extracting descriptions, and determining the number of candidate images by analyzing the layout (rows and cells). Conversely, the latter module aims to associate objects (i.e., candidate images) with their respective cells in the challenge. The authors highlighted their use of a fast object detector, YOLOv3, to meet the time constraints inherent in solving reCAPTCHA challenges. They trained two YOLOv3 variants using two datasets: the first derived from MS COCO [84], containing 8 image classes commonly found in reCAPTCHA tests, and the second dataset compiled from images sourced from Flickr, Google, and Bing, along with an additional 2,100 images extracted from various reCAPTCHA challenges. The solver achieved an accuracy of 83.25% in breaking real-world reCAPTCHA v2 challenges within an average time of 19.93 seconds, inclusive of network-related delays.

Table 1 offers a comprehensive summary of the methods discussed, detailing the deep learning models utilized, training and inference times, accuracy rates, and the datasets on which they were tested.

2) ATTACKS ON OTHER IMAGE CAPTCHAS

Aside from Google's reCAPTCHA, there are numerous other significant and extensively employed image CAPTCHAs. While they share some conceptual and modal similarities, each presents its own unique set of challenges to be solved. Consequently, there has been a notable uptick in the development of tailored attacks aimed at circumventing these modern CAPTCHAs. This section delves into the discussion of state-of-the-art attacks of this nature.

The CAPTCHA used in China's railroad system, akin to Google's reCAPTCHA v2, is the focus of investigation in [85]. In contrast to reCAPTCHA, this system incorporates two protective measures designed to thwart automated solutions. Firstly, rather than providing plain text instructions, the solving process is outlined through noisy character images. Secondly, the challenge images (i.e., candidate images) are deliberately presented in low resolution, adding complexity to the automatic recognition process. The authors propose a solution involving clustering of challenge images and learning associations between text descriptions and image clusters. Utilizing co-occurrence information, they construct a graph to capture relationships among images and between images and descriptive text. Deep convolutional neural networks are employed to encode image descriptions, trained on a set of 230 randomly selected Chinese phrases, creating a probability distribution in a 230-dimensional latent feature space. For image recognition, a weighted graph containing 3.5 million images sourced from real-world China's railroad CAPTCHA challenges is constructed in which nodes represent images, and edges signify similarity, calculated using CaffeNet, and co-occurrence records. To facilitate efficient comparison and matching, perceptual hashing [86] is utilized, assigning similar values to similar images. To solve a CAPTCHA, the proposed method compares eight challenge images with vertices in the graph, processing the images to assign appropriate tags. Subsequently, the system selects four images that best match the test description based on a predetermined threshold. If no image meets the criteria, a single image, deemed the best possible match, is selected as the response. Through testing with over 7,000 manually labeled descriptions, the proposed strategy achieves an 80% accuracy rate in interpreting descriptions, with a CAPTCHA breaking success rate of 77%.

SelAttack, introduced in [59], was also utilized to target China's railway CAPTCHA, employing two CNN models: one for breaking image-based challenges and the other for interpreting text descriptions. The second model was trained on a custom dataset of CAPTCHA descriptions gathered from real-world instances of the railway's online system. Experimental results reported a success rate of 90% within an acceptable timeframe of less than five seconds.

To enhance the security of modern image-based CAPTCHAs, neural style transfer [87], [88] was adopted in [89] to create Grid-CAPTCHA (refer to Figure 4). The fundamental concept of this approach involves leveraging deep learning, specifically the neural style transfer method, to modify an existing image (referred to as the content image) based on a style image. Essentially, this technique introduces deep noise, as opposed to standard perturbations, to an image to make it difficult for machines to recognize. The appearance of Grid-CAPTCHA resembles that of reCAPTCHA v2, thus the proposed deep learning attack shares similarities with those discussed in the previous section.

The initial step of the conducted attack in this work involves interpreting the CAPTCHA description, which is safeguarded with distortion, rotation, and background noise, using an LSTM network with an attention mechanism. This recurrent neural network was trained on a dataset comprising 48,000 machine-generated descriptions, similar to those presented in China's railway system CAPTCHA. Additionally, an additional 2,000 CAPTCHA samples were procured from the target website to aid in the extraction of candidate images from the CAPTCHA. Feature extraction

TABLE 1. Summary of deep learning methods for breaking reCAPTCHA.

Work	Year	Target	Model	Training Time	Running Time	Accuracy	Dataset
Goodfellow et al., [71]	2014	reCAPTCHA Street View	Customized CNN	6 Days	N/A	96%	SVHN
Zhou et al., [77](#1)	2018	reCAPTCHA (road sign challenges)	Customized CNN (GoogleNet)	N/A	N/A	94.25%	Custom (Google Image Search)
Zhou et al., [77](#2)	2018	reCAPTCHA (road sign challenges)	Customized CNN	N/A	N/A	77.2%	Custom (Google Image Search)
Zhao et al., [59](#1)	2018	reCAPTCHA (2015 version)	Customized CNN	18 Hours	1.26 Seconds	88%	ImageNet + Image search(from Baidu and Google)
Zhao et al., [59](#2)	2018	reCAPTCHA (2018 version)	Customized CNN	8 Hours	4.92 Seconds	79%	ImageNet + Image search(from Baidu and Google)
Wang et al., [80](#1)	2020	reCAPTCHA v2	YOLO v3	12 Hours	N/A	61%	ImageNet + Image search (Google)
Wang et al., [80](#2)	2020	reCAPTCHA v2	YOLO v3	12 Hours	N/A	68%	Custom (reCAPTCHA samples)
Hossen et al., [83](#1)	2020	reCAPTCHA v2	YOLO v3	15 Days (on the first dataset) / 2 Days (on the second dataset)	19.93 Seconds	83.25%	MSCOCO/Custom (Google, Flickr,Bing)

is carried out using SE-ResNext-101 [90], pre-trained on ImageNet. Answers are selected based on the acquired feature vectors. Specifically, by computing the distance and cosine similarity between image pairs, in relation to a pre-specified threshold, the similarity between each challenge image and every image in the customized dataset (gathered from real-world railway CAPTCHAs) is assessed. The most frequent answers (appearing more than five times) in the list of probable answers are selected as the final solution. The attacks successfully deciphered the description with 95% accuracy for China's railway CAPTCHA and nearly 100% accuracy for Grid-CAPTCHA. Finally, the solving technique achieved a 62% success rate (in 0.73 seconds) for the former CAPTCHA, while it did not achieve more than 7% accuracy in 0.8 seconds when targeting the Grid-CAPTCHA.

In another endeavor to exploit neural style transfer, Style Area CAPTCHA (SACaptcha) is proposed in [91]. The underlying concept of this HIP test involves transforming an image into a synthesized one by adding shapes and providing a text description. Users are required to click on the added shapes, i.e., foreground regions/objects, to verify their humanity. To assess the effectiveness of SACaptcha, the authors developed two deep learning-based attacks. The first attack employed an object detection approach using a Faster R-CNN network, trained using CAPTCHA challenge images alongside shape and location information of foreground regions. The solver model was tested on three versions of SACaptcha. The most complex test, consisting of CAPTCHAs with 25 foreground shapes and 11 different styles, was solved with an average time of 0.15 seconds and an accuracy of only 5%. Conversely, the attack on the simplest version, with two foreground shapes and 11 styles, achieved a success rate of 19.9% in an average time of 0.12 seconds. The second attack utilized a fully convolutional semantic segmentation network introduced in [92]. This network was trained with CAPTCHA challenge and mask images, with background pixels and pixels related to the style (foreground objects) distinguished with different colors in the masks. The success rate of the model for the most complex version was 4.5% in an average time of 2.20 seconds, while for the simplest version, it achieved an accuracy of 43.6% in 2.16 seconds on average.

A recent study targeting SACaptcha is introduced in [93]. The devised model is based on Mask R-CNN [94] and, for each candidate object, it outputs a class label, a bounding-box offset, and an object mask. The model was trained on

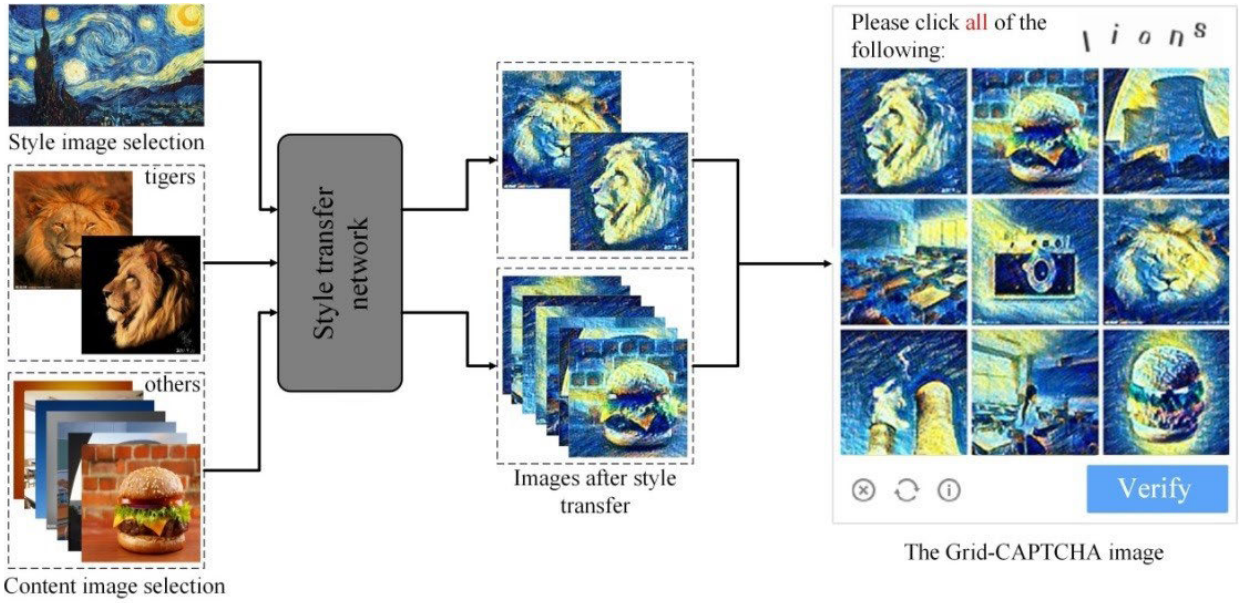


FIGURE 4. The architecture and example of the Grid-CAPTCHA [89].

a custom dataset comprising 300 neural style transferred images generated from 75 content images and five style images. Additionally, the shapes were annotated using the VGG Image Annotator.[3] Experimental results, measured in terms of precision and F1 score, demonstrate the effectiveness of the approach across both employed datasets. Specifically, over the first dataset, the approach achieves a precision of approximately 96.0% and an F1-score of 0.828, while on the second dataset, precision and F1-score are 100% and 0.962, respectively. An important advantage of the proposed attack is its robustness against irregular shapes, enabling it to handle various stylized challenges.

SCAPTCHA [95] is another recently introduced CAPTCHA that utilizes object segments and their related metadata to generate challenges. These object segments, suitably manipulated through rotations and occlusions, are gathered by crawling the web and labeled under classes not common in popular datasets such as ImageNet and MS COCO, thus limiting the effectiveness of pre-trained DNNs for breaking challenges. The manipulated segments are randomly selected and positioned in a noisy background to construct the final CAPTCHA. To provide the description, or the CAPTCHA question, cognitive questions are formed, such as asking users to specify the correct number of a specific object in a challenge. SCAPTCHA can be presented in two configurations, single and triple, as illustrated in Figure 5, wherein users must pass (answer) three challenges (questions).

To assess the robustness of SCAPTCHA against state-of-the-art solvers, an attack utilizing MASK R-CNN, trained on the MS COCO dataset, was conducted on 300 SCAPTCHA instances targeting three common semantic classes: human,

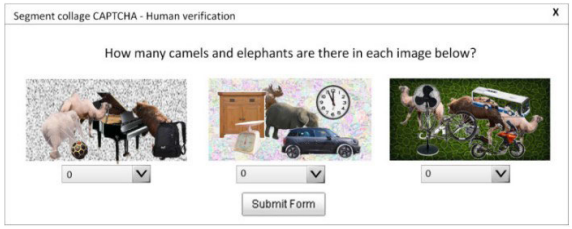


FIGURE 5. A sample SCAPTCHA test with triple challenges [95].

zebra, and elephant. For single SCAPTCHAs, the model achieved success rates of 9% for humans, 20% for zebras, and 15% for elephants. However, in the case of triple challenges, the accuracy dropped significantly to 0.07% for humans, 0.8% for zebras, and 0.33% for elephants.

In order to break hCaptcha,[4] a modern image-based CAPTCHA similar in design and user interface to reCAPTCHA v2, Hossena and Hei [96] employed a standard DNN approach. This approach utilized a pre-trained ResNet-18, trained on ImageNet, with the last classification layer modified to accommodate the nine classes of hCaptcha. The model was evaluated on 270 hCaptcha challenges and achieved an accuracy of 95.93% within an average time of 18.76 seconds.

Visual reasoning-based CAPTCHAs [97], which require users to perform highly cognitive tasks such as selecting the blue letter furthest from the red circle, pose significant challenges for machines. To crack a prominent CAPTCHA of this nature, specifically Tencent's Visual Turing Test (VTT) CAPTCHA, a novel technique is proposed in [98]. In the VTT CAPTCHA, users are tasked with clicking any pixel within the target object. In this approach, each image

[3]https://www.robots.ox.ac.uk/ vgg/software/via/

[4]https://www.hcaptcha.com/

is initially divided into a 14 × 14 grid, with predictions made for each cell regarding whether the center coordinates of the target object lie within it. The proposed model comprises three modules: input, reasoning, and output. The input module serves as a feature extractor, responsible for semantically interpreting challenge instructions through a BiLSTM [99] and extracting visual features using ResNet-50 [100]. The reasoning module, leveraging a modified version of the attention-based MAC network [101], predicts the location information of the answer object by relating the description and reasoning steps. Finally, the output module determines the cell to be clicked by integrating interpreted text information and memory state (containing location information) using a classifier with two fully-connected layers. Training for this attack was conducted on 13,500 VTT CAPTCHA challenges, with a subset of 10,000 instances used for training and validation and test datasets comprising 2,500 and 1,000 instances, respectively. The proposed method achieved an average success rate of 67.3% with a noteworthy processing time of less than 0.05 seconds for VTT CAPTCHA. Additionally, this attack, with an average speed of 0.96 seconds for all tests, was applied to Geetest CAPTCHA with a success rate of 66.7%, to NetEase with a success rate of 77.8%, and achieved a success rate of 86.5% for Dingxiang CAPTCHA.

The authors also proposed an alternative attack strategy specifically targeting other visual reasoning CAPTCHAs, including Geetest, NetEase, and Dingxiang. They utilized a dataset comprising 5,000 instruction-image pairs collected from the respective CAPTCHA instances. The proposed attack framework consists of four main modules: semantic parsing, detection, classification, and integration. The semantic parsing module, as its name suggests, addresses the reasoning tasks required to solve the test. During this step, the textual instructions of the test are parsed using a two-layer LSTM network. The detection module, based on R-CNN, and the classification module, powered by SENet, perform computer vision tasks such as object locating and image-level feature extraction. Finally, the integration module combines semantic information and object-related features to produce the answer to the CAPTCHA test. The remarkable reported success rates for this modular attack are as follows: 90.8% for Geetest, 86.2% for NetEase, and 98.6% for Dingxiang. Additionally, it is noted that the longest operation time recorded for these tests was for Geetest, lasting 10.7 seconds.

Building upon their prior research, the authors conducted further investigation into compromising visual reasoning CAPTCHAs. In addition to the data gathered in their previous study [98], they collected an additional 5,000 prompt-image pairs from the Shumei[5] and Xiaodun[6] CAPTCHA schemes. The success rates of the new attacks on Shumei and Xiaodun CAPTCHAs were reported as 95.9% and 79.2%, respectively. Additionally, the average solving speeds for these CAPTCHAs were reported as 0.79 seconds for Shumei and 0.97 seconds for Xiaodun.

In the pursuit of enhancing visual reasoning CAPTCHAs, a new CAPTCHA scheme called Common Sense CAPTCHA (CsCAPTCHA) [102] is introduced as an additional contribution. This scheme, both in appearance and functionality, bears resemblance to modern image CAPTCHAs. However, users are presented with common-sense challenges and are required to select the appropriate image from a 12-choice answer space. They pass the test upon successfully navigating two consecutive challenges.

To generate the textual questions, keywords are extracted from benchmark word repositories such as WordNet [103], CommonsenseQA [104] and [105]. These keywords are then used to establish relationships and generate statements (textual instructions) using common-sense knowledge, sourced primarily from ConceptNet [106]. The test images are sourced from Baidu and Google search engines, then undergo preprocessing, including rotation and dilation, using the VGG model. To enhance the security of the test images, additional adversarial noise is introduced through the utilization of the SI-NI-FGSM method [107].

A usability comparison reveals its superior human success rate compared to reCAPTCHA v2 (70.7% versus 57.9%), while exhibiting a shorter response time (18.5 seconds) in contrast to reCAPTCHA v2, which averages 25.7 seconds.

To evaluate the resilience of CsCAPTCHA, various attack scenarios were examined. A traditional Brute Force attack yielded a success rate of less than one percent, while the holistic attack introduced in [98] failed to compromise CsCAPTCHA after 1000 attempts.

For a deeper analysis of CsCAPTCHA's security, a modular attack was devised, comprising four primary phases: noun extraction, retrieval of common-sense knowledge, category classification, and text-image matching. CoreNLP [108] was employed to extract nouns from the common-sense sentences (i.e., textual instructions). Subsequently, ConceptNet was queried using the identified nouns to retrieve candidate phrases. Finally, images whose labels (classified by ResNet50 on ImageNet) matched the candidate phrases were considered potential answers. This sophisticated attack achieved a success rate of only 1.1% after 1000 attempts.

The research conducted in [109] thoroughly examines the security and robustness of slider-based (image) CAPTCHAs. In this study, five slider CAPTCHAs (all in Chinese language) are subjected to the attack consists of two main phases: 1) determining the final position of the slider (as completed by humans) and 2) mimicking mouse movements to drag the slider to the determined position. For the first part, Faster R-CNN is utilized to reconstruct the sliding trajectory based on identified key points from the background image (of the challenge). Subsequently, mouse movement simulation is employed to submit responses as genuine human users would. The reported success rates (100% for Taobao, 87.5% for GeeTest, 91% for Netease, 85% for Tencent, and 89.8%

[5]https://www.ishumei.com/trial/captcha.html

[6]https://sec.xiaodun.com/onlineExperience/spatialReasoningSel-ection

for VAPTCHA) underscore the efficacy of the proposed methodology.

Table 2 provides summary of the methods overviewed in this section.

V. DEEP LEARNING FOR PROTECTING IMAGE CAPTCHAS

Although achievements in breaking CAPTCHAs have shown that there is not any unsolvable challenge, for years, many efforts to protect the CAPTCHA from adversarial attacks have also been proposed. The earliest efforts aimed to make image CAPTCHAs as difficult as possible for computer vision attacks by adding visible noises. However, although such an approach was successful in protecting the challenge; the usability was remarkably decreased [110]. To cope with such side effects, several deep learning-based techniques have been investigated to protect the CAPTCHA while at the same time preserving the usability. This research field has tackled both text-based [111], [112], [113], [114], [115] and image-based CAPTCHAs. DeepCAPTCHA [116], introduced in 2017, adds noise to images in a way that it cannot be removed by deep learning methods. It, specifically, proposes the concept of Immutable Adversarial Noise (IAN) driven by two main requirements: 1) being able to fool deep models in at least 98.5%, and 2) not affecting human recognition as well as being computationally efficient. For generating the adversarial examples that preserve the semantic class of the original image, the authors proposed an iterative version of the fast gradient sign method (FGS), i.e. IFGS. IFGS, besides the original image, receives a target label and a confidence level to make sure that the generated adversarial sample is not semantically different from the primary image and at the same time keeps the noise minimal. This process, then, was employed to generate adversarial image challenges for the DeepCAPTCHA. The subject image for noise injection is selected randomly from ImageNet ILSVRC-12 dataset and labeled by using CNN-F [117]. The usability tests show 82.57% success rate for solving the DeepCAPTCHA by human users. The capability of some standard deep models, namely AlexNet [118], CNN-M and CNN-S similar to the network introduced in [119] and OverFeat [120], in recognizing the generated adversarial images were also examined and none of them was able to reach a pre-determined security threshold of 1.5%.

The idea proposed in [121] for generating adversarial image CAPTCHA is to add as much as noise to an image until it remains recognizable by humans. In this regard, they designed an adversarial CAPTCHA generation system, aCAPTCHA. Comprehensive experiments validated the robustness of this technique against ResNet50, GoogleNet, VGG and NetInNet networks.

CAPTURE [122] was introduced as another CAPTCHA scheme based on constrained adversarial perturbation technique. To generate adversarial images, two approaches are proposed: 1) synthesis of unrecognizable images with indirect encoding that employs compositional pattern producing network (CPPN) [123] and, 2) generation of adversarial patches [124]. Although the conducted experiments demonstrated the robustness of the two strategies against classifiers trained on VGG16 & VGG19 [125], ResNet-50, InceptionV3 [126], Xception [127] and MobileNet [100], the significant changes introduced on the source image makes the approach weak against attacks based on either detecting unusual objects or saliency detection (see Figure 6).

FIGURE 6. Example of adversarial patched CAPTCHA [122] with apparent visual noise.

GARD-CAPTCHA [128] protects image challenges by adding noises that in fact are some information-rich pixels from one to several images and injects them into the original image to deceive classifiers. The robustness of this method was evaluated (and confirmed) by models trained on VGG-16, AlexNet and SqueezeNet [129].

To enhance the resilience of the CAPTCHA challenge against deep learning attacks, TICS [130] proposes a text-image CAPTCHA incorporating semantic (cognitive) challenges. The test generation for TICS involves two primary steps. Firstly, it generates an image using semantic GAN based on an initial image and a textual description. Secondly, leveraging a multi-condition GAN network (MC-GAN [131]), it produces a syntactic image based on the initial text description overlaid on the background of the source image. According to the paper, CNN-based classification struggled to overcome this challenge.

A frequently successful approach for circumventing a CAPTCHA scheme involves training a model with real-world images (challenges) sourced from the targeted system. To counter such exploitation, researchers in [132] have introduced a GAN-based image generation model to continually produce fresh images, bolstering the CAPTCHA system's resistance to attacks. Additionally, addressing the critical need for real-time generation of image tests, which occurs when high-quality artificial images are synthesizing using GANs, they employed delay-aware Lyapunov-control-based optimization to preserve the stability the system. This

TABLE 2. Summary of deep learning methods for breaking modern image CAPTCHAs.

Work	Year	Target	Model	Training Time	Running Time	Accuracy	Dataset
D'Souza et al., [69]	2012	AVATAR CAPTCHA	CNN	N/A	N/A	99%	Avatar CAPTCHA dataset
Ya et al., [85]	2017	China's Railway CAPTCHA	DCNN (LeNet), CaffeNet	N/A	2 Seconds	77%	Custom (Target CAPTCHA samples)
Zhao et al., [59] (# 3)	2018	China's Railway CAPTCHA	CNN	93 Hours	4.4 Seconds	90%	ImageNet, Image search (from Baidu and Google)
Cheng et al., [89](# 1)	2018	China's Railway CAPTCHA	SE-ResNeXt-50	N/A	0.73 Seconds	62%	Custom (Target CAPTCHA samples)
Cheng et al., [89](# 2)	2018	Grid-CAPTCHA	SE-ResNeXt-50	N/A	0.8 Seconds	7%	Custom (Grid-CAPTCHA Icons)
Tang et al., [91](Object Detection Attack)	2018	SACaptcha	Faster R-CNN	N/A	0.15 Seconds	5%	Custom (Target CAPTCHA samples)
Tang et al., [91](Pixel-level segmentation attack)	2018	SACaptcha	FCN	N/A	2.20 Seconds	4.5%	Custom (Target CAPTCHA samples)
Nguyen et al., [95]	2020	SACaptcha	Mask R-CNN	N/A	N/A	20% (Single Challenge) 8% (Triple Challenges)	MS-COCO
Rathor et al., [93]	2021	SACaptcha	Mask R-CNN	N/A	N/A	N/A	Custom
Hossen and Hei, [96]	2021	hCaptcha	ResNet-18	143 Minutes	18.76 Seconds	95.93%	ImageNet
Gao et al., [98]	2021	VTT CAPTCHA	MAC, ResNet-50	N/A	0.05 Seconds	67.3%	Custom (Target CAPTCHA samples)
Chang et al., [109]	2024	TaoBao	N/A	N/A	3.78 Seconds	100%	N/A
Chang et al., [109]	2024	Geetest	Faster R-CNN	N/A	10.16 Seconds	87.5%	Custom (Target CAPTCHA samples)
Chang et al., [109]	2024	Netease	Faster R-CNN	N/A	7.29 Seconds	91%	Custom (Target CAPTCHA samples)
Chang et al., [109]	2024	Tencent	Faster R-CNN	N/A	8.3 Seconds	88%	Custom (Target CAPTCHA samples)
Chang et al., [109]	2024	VAPTCHA	Faster R-CNN	N/A	8.72 Seconds	89.8%	Custom (Target CAPTCHA samples)

yields through selecting an optimal GAN model to generate high-quality images while caring the time sensitivity of the process.

The fusion of neural style transfer with adversarial examples to safeguard text- and image-based CAPTCHAs is presented in [133]. The methodology is outlined as follows: initially, neural style transfer intricately enhances the target (output) image, followed by the injection of additional adversarial noise into the stylized image, thus fortifying its defense against computer vision attacks. For the former step, the Fast Style Transfer Network [134] is used, and the Fast Gradient Sign Method (FGSM) [135] is employed to generate adversarial examples. Random characters for generating text images are randomly chosen from the EMNIST [136] dataset, while images (along with style images) are sourced from ImageNet. This hybrid approach reported a commendable resistance against image recognition attacks executed by ResNet-101 and VGG-16. For instance, a ResNet-driven attack achieved a 67.7% success rate when targeting stylized tests and a 45.7% success rate in circumventing adversarial tests (generated by the same network), while the success rate for stylized adversarial examples was 35.7%. Text-based examples exhibit an even higher resilience against attacks leveraging ResNet-50 and LeNet-5 [137] networks. Specifically, a LeNet-based attack on stylized, adversarial, and stylized adversarial examples yielded success rates of 37.6%, 17.3%, and 7.8%, respectively. The authors originally introduced the foundational concept behind this work to present zxCAPTCHA [138], which is an image-based challenge designed to integrate text, image, and cognitive elements aiming to generate a stylized adversarial CPATCHA test.

The CAPTCHA challenge outlined in [102] presents an innovative approach by reversing the typical recognition process found in current CAPTCHAs based on Style Transfer. In this new approach, termed Style Matching CAPTCHA by the authors, users are tasked with matching the styles applied to a given query (content) image with a series of candidate style images.

A. CONTEXT-AWARE PROTECTION

To enhance the security of modern image CAPTCHAs and effectively combat potential attacks, a context-aware approach is imperative. Addressing the vulnerabilities that will be discussed in the attack workflow necessitates a comprehensive strategy that acknowledges two key interrelated weaknesses. Consequently, implementing separate protective measures becomes crucial to bolster the resilience of the CAPTCHA system. What underscores the importance of this approach is the intrinsic variance between safeguarding text-based instructions and challenge images, as well as their symbiotic role in the CAPTCHA-solving process. Neglecting to reinforce the security of the textual component could inadvertently create a vulnerability exploited to bypass image CAPTCHAs.

The methodologies reviewed in Section IV share a common conceptual framework, as depicted in Figure 7, illustrating an attack on a contemporary (selection-based) image CAPTCHA. In this architecture, a state-of-the-art solver initially dissects the CAPTCHA, conducting browser-based operations to handle the textual instructions and images separately. Depending on the complexity of the instructions, tasks such as denoising, character recognition, and image interpretation are performed. Simultaneously, deep learning algorithms are employed to classify the challenge images, with the accuracy of this classification profoundly influencing the overall effectiveness of the process. Subsequently, a matching network establishes the correlation between the textual instructions and the classified/annotated images, often utilizing scores assigned to each pair to select the final answer(s). Finally, browser-based operations are utilized to submit the answer(s), with incorrect submissions prompting a repetition of the previous steps.

As previously discussed, breaking image CAPTCHAs involves two primary processes. It is essential to note that the initial step involves acquiring real-world samples of the target CAPTCHA, in addition to utilizing extensive and diverse image datasets for training models. The first process is addressing the textual description outlining the steps to solve a specific challenge. These instructions can be safeguarded through various means, such as using character images or introducing noise and geometric modifications (similar to the approach taken by China's railroad CAPTCHA). However, these measures can be circumvented by computer vision techniques, particularly those based on deep learning. The challenge intensifies when the comprehension of the instructions necessitates reasoning and interpretation, as misinterpretations can significantly impede progress. In such cases, advanced natural language processing techniques and large language models [139], can provide effective support.

However, some CAPTCHAs do not provide explicit textual information, requiring agents (machines) to solve the challenge by matching supplied guide images with candidates (as illustrated in Figure 8). This scenario demands an additional image recognition operation, particularly on small-size and low-quality images. Once the solution methodology is understood, the second main step involves employing image classification models to identify image candidates relevant to the description. Subsequently, the best match(es), determined by criteria such as similarity measures, are selected as the answer(s).

To devise a robust strategy for safeguarding CAPTCHAs against deep learning-driven attacks, it is essential to embrace a context-aware approach. As depicted in Figure 7, the two primary components targeted in breaking image CAPTCHAs are the textual description and the challenge images. Consequently, each of these elements necessitates tailored protection measures based on their unique characteristics.

Employing traditional techniques like adding noise, akin to classic text-based CAPTCHAs, is no longer sufficient to secure the textual description in image CAPTCHAs.

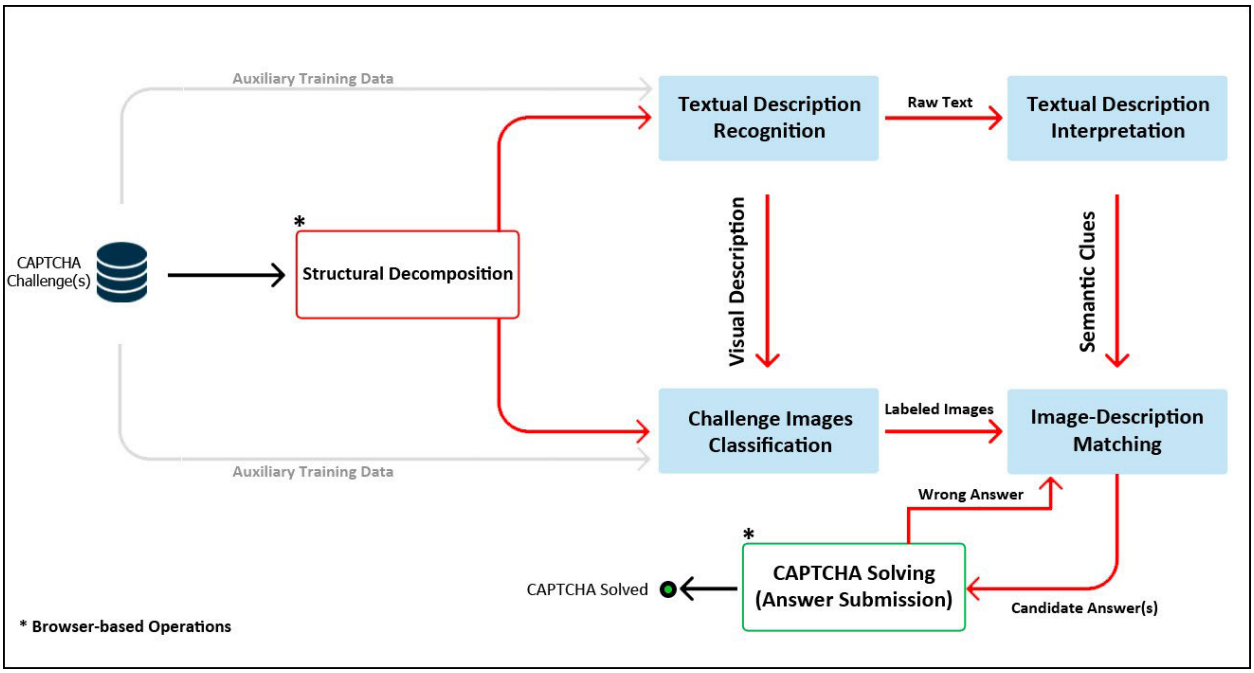


FIGURE 7. Conceptual framework of breaking image-based CAPTCHAs.

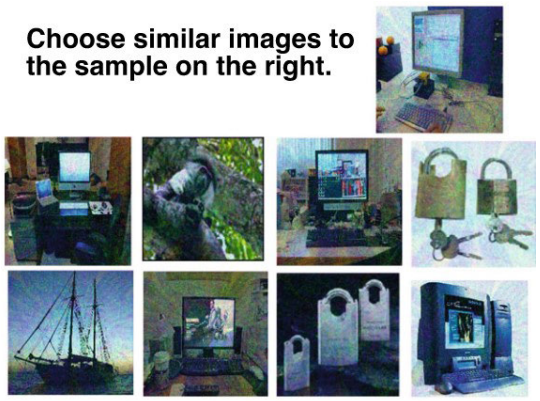


FIGURE 8. An image CAPTCHA [128] with less-informative text description.

Instead, leveraging adversarial text generation methods, capable of deceiving deep learning-based classifiers, proves to be a promising tactic [140], [141]. These methods adding a security layer on crucial keywords within the textual information by subtly altering characters, punctuation marks, or sentence structures to render them readable for humans while remaining imperceptible to machines.

However, while generating adversarial images is a common strategy to impede machine-based image recognition, certain drawbacks plague some of these methods. Particularly, generating excessively noisy images or those with conspicuous perturbations not only compromises user-friendliness but also renders them susceptible to detection by simple computer vision attacks. To mitigate such issues, noise should be incorporated in a manner that minimizes visibility. For instance, techniques like SAI-FGM [142] enhance sample security against attacks while simultaneously reducing noise. By applying perturbations at the superpixel level, these approaches offer more effective solutions compared to pixel-wise methods. Additionally, strategies such as modification of semantic information [143], targeting semantic similarity [144], leveraging spatial constraints [145], and employing adversarial color transformation [146] can be deployed.

Moreover, incorporating adversarial samples within training datasets can disrupt the learning process for attackers, serving as an additional layer of protection [83]. This proactive measure undermines the efficacy of attackers' methods trained on real-world challenge images acquired from the target service, thereby enhancing the overall security posture.

VI. DISCUSSIONS AND FUTURE WORKS

Breaking modern image-based CAPTCHAs involves two primary, interconnected workflows: interacting with the CAPTCHA and solving it. In the former, the process begins with acquiring the CAPTCHA, understanding how the challenge is presented, generating a solution, and ultimately submitting the answer back to the CAPTCHA system. These steps must be executed carefully to mimic actions of a genuine user, as failure to do so could trigger the renewal of challenges, necessitating a restart of the entire process. The latter workflow encompasses the application of computer vision and deep learning techniques, tasked with classifying challenge images and determining potential answers based on the provided instructions.

Specifically, to equip deep learning-based computer vision techniques with the necessary challenge images, precise identification of the grid-like structure of the CAPTCHA is essential to extract thumbnail-sized images for model input. Conversely, both answer recognition and submission must be completed within a restricted time frame to circumvent time-based security measures.

Considering the points discussed above, the notion of crafting generic image CAPTCHA solvers appears not only impractical but potentially unattainable. Many of the existing studies outlined in this review primarily focus on breaking the challenge itself, without delving into the practical implications of such attacks in real-world scenarios. In essence, transforming the process of image recognition and comprehension into a universally applicable solving strategy necessitates accounting for various influential factors. These factors include network delays, as highlighted in studies such as [59] and [111], encountering multiple challenges to bypass the CAPTCHA, and adapting to changes in challenge types.

For instance, in the case of reCAPTCHA v2, it is common for the challenge type to shift after an unsuccessful attempt, transitioning from tasks like identifying a specific object to selecting a cell associated with a particular concept. The approach outlined in [80] serves as a notable example of devising a comprehensive strategy to address such scenarios.

Another crucial factor influencing the automated solving of image CAPTCHAs is the availability of suitable datasets. Typically, these datasets are compiled by gathering real-world samples of the target CAPTCHA. However, as noted in studies like [59] and [96], leveraging existing extensive datasets like ImageNet can significantly enhance the learning process. A reliable strategy involves utilizing a hybrid dataset, which combines a general-purpose dataset with a tailored one containing CAPTCHA samples. This hybrid approach compensates for any deficiencies in category coverage within the broader dataset.

However, it is important to acknowledge that when CAPTCHA systems employ generative models to generate challenge images [132] or when these challenges are fortified with adversarial noise and examples [83], the reliability of collecting datasets of this nature comes into question.

Thanks to advancements in computer vision techniques, several innovative methodologies in the field of deep leaning-based computer vision have emerged that can be employed for solving image CAPTCHAs in more efficient and creative ways. For example, as the latest breakthrough in the field,image foundation models [147] can introduce unprecedented functionality in decoding both textual and visual challenges within CAPTCHA tests. Further exploration of similar approaches is discussed below.

A. SALIENT OBJECT DETECTION

Salient object detection methods offer a potent solution for pinpointing the most significant object(s) within an image (refer to Figure 9). This capability streamlines the process of identifying candidate answers. Consider a scenario where multiple objects populate the scene; assessing image similarity becomes notably intricate. However, by pinpointing salient objects within each challenge image, the annotation [148] and retrieval [149] processes become more precise. Furthermore, the integration of salient object detection with weakly supervised object detection [150] and semantic segmentation [151] can yield a substantial boost in the performance and accuracy of the solver model. These techniques prove particularly advantageous when the CAPTCHA system conveys its instructions through images rather than text.

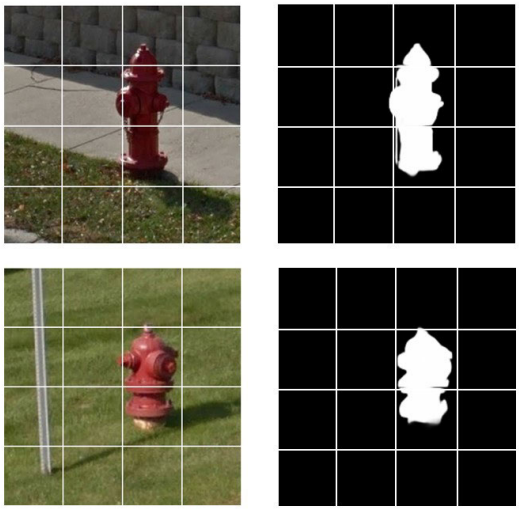


FIGURE 9. An illustration of the utility of salient object detection in overcoming CAPTCHA challenges. In these instances, objects identified as significant through salient object detection methods facilitate the efficient resolution of CAPTCHAs.

B. IMAGE-TEXT MATCHING

The challenge of identifying one or more images that semantically align with a given text description can be reframed as an image-text matching problem. By harnessing cutting-edge techniques in this domain, as exemplified by [152], [153], [154], [155], and [156], effective selection of matching candidates becomes feasible following the interpretation of the CAPTCHA description (see Figure 10). However, when prior knowledge about the target CAPTCHA suggests that the candidate images belong to a specific category, such as those related to traffic scenes, contextual information from both text and images becomes pivotal [157].

C. IMAGE SIMILARITY COMPUTATION

A promising avenue for future research involves employing similarity computation techniques [158], [159], [160] to align visual descriptions with candidate images. The fundamental concept is to discern analogous structures and/or content within images. Utilizing deep learning-based similarity assessment methods further streamlines the matching process by correlating candidate images with labeled instances in a dataset. As demonstrated in [161], integrating domain-specific knowledge and requirements significantly

FIGURE 10. An example of matching a text query with image candidates generated by the method proposed in [155] (image in the green box selected as the correct answer).

enhances accuracy, particularly in specialized use cases. Once again, the example of road sign and traffic scene CAPTCHAs underscores the potential for training case-specific algorithms to detect subtle similarities among images.

From another perspective, to enhance the security of CAPTCHA systems and protect them from being automatically cracked, the following measures can be considered:

D. SEMANTIC VISUAL CHALLENGES

To complicate the interpretation of visual challenges by advanced computer vision algorithms, adding semantic layers can be highly effective. For instance, challenges that require clicking on an object of a specific color in a particular location and with a specific feature could present a significant difficulty for machines. Studies on machine-driven attacks on visual reasoning CAPTCHAs [97] demonstrate that semantically designed visual tests can significantly resist advanced attacks.

E. USER-FRIENDLY ADVERSARIAL PROTECTION

As mentioned earlier, using adversarial patches may not be effective for protecting CAPTCHA tests due to their conspicuous appearance. However, recent advancements [162], [163], [164] in this field have introduced effective, user-friendly solutions. These solutions involve adversarial effects that are subtle enough to be imperceptible to humans, ensuring an unaltered visual experience, while still being difficult for machines to detect and bypass. These adversarial techniques can mislead algorithms when they attempt to recognize objects within the CAPTCHA test.

VII. CONCLUSION

Image-based CAPTCHAs have gained prominence in recent years among various types of CAPTCHAs. This surge in popularity can be attributed to significant strides made in circumventing traditional text-based CAPTCHAs. As a result, with machines still lagging behind humans in solving cognitive image-oriented authentication tests, such challenges appear effective in thwarting bots from mimicking genuine user behavior. However, researchers and adversarial parties alike have begun leveraging deep learning techniques to break modern image CAPTCHAs, driven by diverse motivations ranging from assessing method robustness to identifying vulnerabilities in current CAPTCHA implementations. In this study, we provide a comprehensive analysis aimed at researchers and practitioners, focusing specifically on recent advancements in the field, with an emphasis on the role of deep learning in fortifying image CAPTCHAs against state-of-the-art attacks.

REFERENCES

[1] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn. Adv. Cryptol. (EUROCRYPT)*, Warsaw, Poland. Berlin, Germany: Springer, 2003, pp. 294–311.

[2] BuiltWith. *CAPTCHA Usage Distribution in the Top 1 Million Sites*. Accessed: Oct. 15, 2023. [Online]. Available: https://trends.builtwith.com/widgets/captcha

[3] J. Chen, X. Luo, Y. Guo, Y. Zhang, and D. Gong, "A survey on breaking technique of text-based CAPTCHA," *Secur. Commun. Netw.*, vol. 2017, no. 1, 2017, Art. no. 6898617.

[4] P. Wang, H. Gao, X. Guo, C. Xiao, F. Qi, and Z. Yan, "An experimental investigation of text-based CAPTCHA attacks and their robustness," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–38, Sep. 2023.

[5] S. Stabinger, A. Rodríguez-Sánchez, and J. Piater, "25 years of CNNs: Can we compare to human abstraction capabilities?" in *Proc. 25th Int. Conf. Artif. Neural Netw. Artif. Neural Netw. Mach. Learn. (ICANN)*, Barcelona, Spain. Cham, Switzerland: Springer, 2016, pp. 380–387.

[6] F. Fleuret, T. Li, C. Dubout, E. K. Wampler, S. Yantis, and D. Geman, "Comparing machines and humans on a visual categorization test," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 43, pp. 17621–17625, Oct. 2011.

[7] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *Proc. 26th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2017, pp. 1–7.

[8] J. Yan, "Bot, cyborg and automated Turing test," in *Proc. 14th Int. Workshop Secur. Protocols* (Lecture Notes in Computer Science), vol. 5087, B. Christianson, B. Crispo, J. A. Malcolm, and M. Roe, Eds. Cambridge, U.K.: Springer, 2006, pp. 190–197, doi: 10.1007/978-3-642-04904-0_26.

[9] J. Yan and A. S. El Ahmad, "A low-cost attack on a Microsoft captcha," in *Proc. 15th ACM Conf. Comput. Commun. Secur.*, Oct. 2008, pp. 543–554.

[10] S. Li, S. A. H. Shah, M. A. U. Khan, S. A. Khayam, A.-R. Sadeghi, and R. Schmitz, "Breaking e-banking CAPTCHAs," in *Proc. 26th Annu. Comput. Secur. Appl. Conf.*, Dec. 2010, pp. 171–180.

[11] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, A. Lavin, and D. S. Phoenix, "A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs," *Science*, vol. 358, no. 6368, Dec. 2017, Art. no. eaag2612.

[12] Y. Wang, Y. Wei, M. Zhang, Y. Liu, and B. Wang, "Make complex CAPTCHAs simple: A fast text captcha solver based on a small number of samples," *Inf. Sci.*, vol. 578, pp. 181–194, Nov. 2021.

[13] R. S. Bhowmick, I. Ganguli, J. Paul, and J. Sil, "Effectiveness of decoder transformer network in breaking low-resource real-time text CAPTCHA system," in *Proc. Int. Conf. Cyberworlds (CW)*, Sep. 2021, pp. 287–290.

[14] O. Bostik, K. Horak, L. Kratochvila, T. Zemcik, and S. Bilik, "Semi-supervised deep learning approach to break common CAPTCHAs," *Neural Comput. Appl.*, vol. 33, no. 20, pp. 13333–13343, Oct. 2021.

[15] Y. Shu and Y. Xu, "End-to-end CAPTCHA recognition using deep CNN-RNN network," in *Proc. IEEE 3rd Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, Oct. 2019, pp. 54–58.

[16] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[17] F.-L. Du, J.-X. Li, Z. Yang, P. Chen, B. Wang, and J. Zhang, "CAPTCHA recognition based on faster R-CNN," in *Proc. 13th Int. Conf. Intell. Comput. Theories Appl. (ICIC)*. Liverpool, U.K.: Springer, 2017, pp. 597–605.

[18] C. Duan, R. Zhang, and K. Qing, "Feature refine network for text-based captcha recognition," in *Proc. 10th Int. Conf. Image Graph. (ICIG)*, Beijing, China. Cham, Switzerland: Springer, 2019, pp. 64–73.

[19] J. Chen, X. Luo, L. Zhu, Q. Zhang, and Y. Gan, "Handwritten CAPTCHA recognizer: A text CAPTCHA breaking method based on style transfer network," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13025–13043, Apr. 2023.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, Y. Bengio and Y. LeCun, Eds. 2016, pp. 1–16.

[22] J. Chang and Y. Gu, "Chinese typography transfer," 2017, *arXiv:1707.04904*.

[23] I. G. Mocanu, Z. Yang, and V. Belle, "Breaking CAPTCHA with capsule networks," *Neural Netw.*, vol. 154, pp. 246–254, Oct. 2022, doi: 10.1016/j.neunet.2022.06.041.

[24] M. Kumar, M. K. Jindal, and M. Kumar, "A novel attack on monochrome and greyscale Devanagari CAPTCHAs," *Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 4, pp. 1–30, 2021.

[25] Y. Jia, W. Fan, C. Zhao, and J. Han, "An approach for Chinese character CAPTCHA recognition using CNN," in *Proc. J. Phys., Conf.*, 2018, vol. 1087, no. 2, Art. no. 022015.

[26] X. Wu, S. Dai, Y. Guo, and H. Fujita, "A machine learning attack against variable-length Chinese character CAPTCHAs," *Appl. Intell.*, vol. 49, pp. 1548–1565, Nov. 2019.

[27] Z. Chen, W. Ma, N. Xu, C. Ji, and Y. Zhang, "SiameseCCR: A novel method for one-shot and few-shot Chinese CAPTCHA recognition using deep Siamese network," *IET Image Process.*, vol. 14, no. 12, pp. 2855–2859, Oct. 2020.

[28] G. Goswami, R. Singh, M. Vatsa, B. Powell, and A. Noore, "Face recognition CAPTCHA," in *Proc. IEEE 5th Int. Conf. Biometrics: Theory, Appl. Syst. (BTAS)*, Arlington, VA, USA, Sep. 2012, pp. 412–417, doi: 10.1109/BTAS.2012.6374608.

[29] J. Kim, S. Kim, J. Yang, J.-H. Ryu, and K. Wohn, "FaceCAPTCHA: A CAPTCHA that identifies the gender of face images unrecognized by existing gender classifiers," *Multimedia Tools Appl.*, vol. 72, no. 2, pp. 1215–1237, Sep. 2014.

[30] S. Kwon and S. Cha, "CAPTCHA-based image annotation," *Inf. Process. Lett.*, vol. 128, pp. 27–31, Dec. 2017.

[31] P. Faymonville, K. Wang, J. Miller, and S. Belongie, "CAPTCHA-based image labeling on the soylent grid," in *Proc. ACM SIGKDD Workshop Human Comput.*, Jun. 2009, pp. 46–49.

[32] J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: A CAPTCHA that exploits interest-aligned manual image categorization," in *Proc. CCS*, vol. 7, 2007, pp. 366–374.

[33] H. Hajjdiab, M. Ghazal, and A. Khalil, "Random image matching CAPTCHA system," *ELCVIA: Electron. Lett. Comput. Vis. Image Anal.*, vol. 16, no. 3, p. 0001, 2017.

[34] H. Gao, D. Yao, H. Liu, X. Liu, and L. Wang, "A novel image based CAPTCHA using jigsaw puzzle," in *Proc. 13th IEEE Int. Conf. Comput. Sci. Eng.*, Dec. 2010, pp. 351–356.

[35] Y. Zhang, H. Gao, G. Pei, S. Luo, G. Chang, and N. Cheng, "A survey of research on CAPTCHA designing and breaking techniques," in *Proc. 18th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun./13th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, Aug. 2019, pp. 75–84.

[36] X. Xu, L. Liu, and B. Li, "A survey of CAPTCHA technologies to distinguish between human and computer," *Neurocomputing*, vol. 408, pp. 292–307, Sep. 2020.

[37] M. Guerar, L. Verderame, M. Migliardi, F. Palmieri, and A. Merlo, "Gotta CAPTCHA'Em all: A survey of 20 Years of the human-or-computer Dilemma," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–33, 2021.

[38] M. Kumar, M. Jindal, and M. Kumar, "A systematic survey on CAPTCHA recognition: Types, creation and breaking techniques," *Arch. Comput. Methods Eng.*, vol. 29, no. 2, pp. 1107–1136, 2022.

[39] N. Tariq, F. A. Khan, S. A. Moqurrab, and G. Srivastava, "CAPTCHA types and breaking techniques: Design issues, challenges, and future research directions," 2023, *arXiv:2307.10239*.

[40] E. Bursztein and S. Bethard, "Decaptcha: Breaking 75% of eBay audio CAPTCHAs," in *Proc. 3rd USENIX Workshop Offensive Technol. (WOOT)*, D. Boneh and A. Sotirov, Eds. Montreal, QC, Canada: USENIX Association, 2009, pp. 1–7. [Online]. Available: https://www.usenix.org/legacy/events/woot09/tech/full_papers/bursztein.pdf

[41] W. Aiken and H. Kim, "POSTER: DeepCRACk: Using deep learning to automatically crack audio CAPTCHAs," in *Proc. Asia Conf. Comput. Commun. Secur.*, 2018, pp. 797–799.

[42] J. Zhang, X. Hei, and Z. Wang, "Typer vs. CAPTCHA: Private information based CAPTCHA to defend against crowdsourcing human cheating," 2019, *arXiv:1904.12542*.

[43] Y. Rui and Z. Liu, "ARTiFACIAL: Automated reverse Turing test using facial features," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 295–298.

[44] A. J. Colmenarez and T. S. Huang, "Face detection with information-based maximum discrimination," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 782–787.

[45] M.-H. Yang, D. Roth, and N. Ahuja, "A SNoW-based face detector," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1–7.

[46] Z. Zhang, L. Zhu, S. Z. Li, and H. Zhang, "Real-time multi-view face detection," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 149–154.

[47] Q. Li, "A computer vision attack on the ARTiFACIAL CAPTCHA," *Multimedia Tools Appl.*, vol. 74, no. 13, pp. 4583–4597, Jul. 2015.

[48] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, Marseille, France. Berlin, Germany: Springer, 2008, pp. 72–85.

[49] P. Golle, "Machine learning attacks against the Asirra CAPTCHA," in *Proc. 15th ACM Conf. Comput. Commun. Secur.*, 2008, pp. 535–542.

[50] C. Javier Hernandez-Castro, A. Ribagorda, and Y. Saez, "Side-channel attack on labeling CAPTCHAs," 2009, *arXiv:0908.1185*.

[51] J. Walker. *A Pseudorandom Number Sequence Test Program*. Accessed: Oct. 20, 2023. [Online]. Available: https://www.fourmilab.ch/random/

[52] J. Friedman, R. Tibshirani, and T. Hastie, "Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors)," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, Apr. 2000.

[53] D. Lorenzi, J. Vaidya, E. Uzun, S. Sural, and V. Atluri, "Attacking image based CAPTCHAs using image recognition techniques," in *Proc. Int. Conf. Inf. Syst. Secur.* Berlin, Germany: Springer, 2012, pp. 327–342.

[54] R. Datta, J. Li, and J. Z. Wang, "IMAGINATION: A robust image-based CAPTCHA generation system," in *Proc. 13th Annu. ACM International Conf. Multimedia*, 2005, pp. 331–334.

[55] B. B. Zhu, J. Yan, Q. Li, C. Yang, J. Liu, N. Xu, M. Yi, and K. Cai, "Attacks and design of image recognition CAPTCHAs," in *Proc. 17th ACM Conf. Comput. Commun. Secur.*, Oct. 2010, pp. 187–200.

[56] H. Gao, L. Lei, X. Zhou, J. Li, and X. Liu, "The robustness of face-based CAPTCHAs," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.; Ubiquitous Comput. Commun.; Dependable, Autonomic Secure Comput.; Pervasive Intell. Comput.*, Oct. 2015, pp. 2248–2255.

[57] G. Goswami, B. M. Powell, M. Vatsa, R. Singh, and A. Noore, "FaceDCAPTCHA: Face detection based color image CAPTCHA," *Future Gener. Comput. Syst.*, vol. 31, pp. 59–68, Feb. 2014.

[58] G. Goswami, B. M. Powell, M. Vatsa, R. Singh, and A. Noore, "FR-CAPTCHA: CAPTCHA based on recognizing human faces," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e91708.

[59] B. Zhao, H. Weng, S. Ji, J. Chen, T. Wang, Q. He, and R. Beyah, "Towards evaluating the security of real-world deployed image CAPTCHAs," in *Proc. 11th ACM Workshop Artif. Intell. Secur.*, 2018, pp. 85–96.

[60] C. J. Hernández-Castro, M. D. R-Moreno, and D. F. Barrero, "Using JPEG to measure image continuity and break capy and other puzzle CAPTCHAs," *IEEE Internet Comput.*, vol. 19, no. 6, pp. 46–53, Nov. 2015.

[61] C. J. Hernández-Castro, M. D. R-Moreno, D. F. Barrero, and S. Gibson, "Using machine learning to identify common flaws in CAPTCHA design: FunCAPTCHA case analysis," *Comput. Secur.*, vol. 70, pp. 744–756, Sep. 2017.

[62] M. Conti, C. Guarisco, and R. Spolaor, "CAPTCHaStar! A novel CAPTCHA based on interactive shape discovery," in *Proc. 14th Int. Conf. Appl. Cryptogr. Netw. Secur. (ACNS)*. Guildford, U.K.: Springer, 2016, pp. 611–628.

[63] C. Hernández-Castro, D. F. Barrero, and M. D. R-Moreno, "Breaking CaptchaStar using the BASECASS methodology," *ACM Trans. Internet Technol.*, vol. 23, no. 1, pp. 1–12, 2023.

[64] C. J. Hernández-Castro, D. F. Barrero, and M. D. R-Moreno, "BASECASS: A methodology for CAPTCHAs security assurance," *J. Inf. Secur. Appl.*, vol. 63, Dec. 2021, Art. no. 103018.

[65] T. Gougeon and P. Lacharme, "How to break CaptchaStar," in *Proc. ICISSP*, 2018, pp. 41–51.

[66] U. Ferraro Petrillo, G. Mastroianni, and I. Visconti, "The design and implementation of a secure CAPTCHA against man-in-the-middle attacks," *Secur. Commun. Netw.*, vol. 7, no. 8, pp. 1199–1209, Aug. 2014.

[67] T.-E. Wei, A. B. Jeng, and H.-M. Lee, "GeoCAPTCHA—A novel personalized CAPTCHA using geographic concept to defend against 3rd party human attack," in *Proc. IEEE 31st Int. Perform. Comput. Commun. Conf. (IPCCC)*, Dec. 2012, pp. 392–399.

[68] L. Clark. *Death to CAPTCHA! Google Wants to Make Them Invisible Using AI*. Accessed: Oct. 20, 2023. [Online]. Available: https://www.wired.co.uk/article/google-wants-to-make-captcha-completely-invisible

[69] D. D'Souza, P. C. Polina, and R. V. Yampolskiy, "Avatar CAPTCHA: Telling computers and humans apart via face classification," in *Proc. IEEE Int. Conf. Electro/Information Technol.*, May 2012, pp. 1–6.

[70] B. Cheung, "Convolutional neural networks applied to human face classification," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, vol. 2, Dec. 2012, pp. 580–583.

[71] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. D. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Y. Bengio and Y. LeCun, Eds., 2014, pp. 1–13.

[72] S. Sivakorn, I. Polakis, and A. D. Keromytis, "I am robot: (Deep) learning to break semantic image CAPTCHAs," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 388–403.

[73] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.

[74] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[75] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[76] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[77] Y. Zhou, Z. Yang, C. Wang, and M. Boutell, "Breaking Google reCAPTCHA V2," *J. Comput. Sci. Colleges*, vol. 34, no. 1, pp. 126–136, 2018.

[78] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[79] F. H. Alqahtani and F. A. Alsulaiman, "Is image-based CAPTCHA secure against attacks based on machine learning? An experimental study," *Comput. Secur.*, vol. 88, Jan. 2020, Art. no. 101635.

[80] D. Wang, M. Moh, and T.-S. Moh, "Using deep learning to solve Google reCAPTCHA v2's image challenges," in *Proc. 14th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2020, pp. 1–5.

[81] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[82] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[83] M. I. Hossen, Y. Tu, M. F. Rabby, M. N. Islam, H. Cao, and X. Hei, "An object detection based solver for Google's image reCAPTCHA v2," in *Proc. 23rd Int. Symp. Res. Attacks, Intrusions Defenses (RAID)*, 2020, pp. 269–284.

[84] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[85] H. Ya, H. Sun, J. Helt, and T. S. Lee, "Learning to associate words and images using a large-scale graph," in *Proc. 14th Conf. Comput. Robot Vis. (CRV)*, May 2017, pp. 16–23.

[86] *Phash: The Open Source Perceptual Hash Library*. Accessed: Oct. 25, 2023. [Online]. Available: http://www.phash.org

[87] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[88] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 694–711.

[89] Z. Cheng, H. Gao, Z. Liu, H. Wu, Y. Zi, and G. Pei, "Image-based CAPTCHAs based on neural style transfer," *IET Inf. Secur.*, vol. 13, no. 6, pp. 519–529, 2019.

[90] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[91] M. Tang, H. Gao, Y. Zhang, Y. Liu, P. Zhang, and P. Wang, "Research on deep learning techniques in breaking text-based CAPTCHAs and designing image-based CAPTCHA," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 10, pp. 2522–2537, Oct. 2018.

[92] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[93] V. S. Rathor, B. Garg, M. Patil, and G. Sharma, "Security analysis of image CAPTCHA using a mask R-CNN-based attack model," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 36, no. 4, pp. 238–247, 2021.

[94] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[95] T. V. Nguyen, Z. Huang, S. Bethini, V. S. P. Ippagunta, and P. H. Phung, "Secure captchas via object segment collages," *IEEE Access*, vol. 8, pp. 84230–84238, 2020.

[96] M. I. Hossen and X. Hei, "A low-cost attack against the hCaptcha system," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2021, pp. 422–431.

[97] H. Wang, F. Zheng, Z. Chen, Y. Lu, J. Gao, and R. Wei, "A CAPTCHA design based on visual reasoning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1967–1971.

[98] Y. Gao, H. Gao, S. Luo, Y. Zi, S. Zhang, W. Mao, P. Wang, Y. Shen, and J. Yan, "Research on the security of visual reasoning CAPTCHA," in *Proc. 30th USENIX Secur. Symp. (USENIX Secur.)*, 2021, pp. 3291–3308.

[99] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[100] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[101] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, May 2018, pp. 1–20. [Online]. Available: https://openreview.net/forum?id=S1Euwz-Rb

[102] P. Ray, A. Bera, D. Giri, and D. Bhattacharjee, "Style matching CAPTCHA: Match neural transferred styles to thwart intelligent attacks," *Multimedia Syst.*, vol. 29, no. 4, pp. 1865–1895, Aug. 2023.

[103] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[104] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," 2018, *arXiv:1811.00937*.

[105] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," 2019, *arXiv:1906.05317*.

[106] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–8.

[107] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," 2019, *arXiv:1908.06281*.

[108] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2014, pp. 55–60.

[109] G. Chang, H. Gao, G. Pei, S. Luo, Y. Zhang, N. Cheng, Y. Tang, and Q. Guo, "The robustness of behavior-verification-based slider CAPTCHAs," *J. Inf. Secur. Appl.*, vol. 81, Mar. 2024, Art. no. 103711.

[110] D. Lorenzi, E. Uzun, J. Vaidya, S. Sural, and V. Atluri, "Enhancing the security of image CAPTCHAs through noise addition," in *Proc. 30th IFIP TC Int. Conf. ICT Syst. Secur. Privacy Protection (SEC)*. Hamburg, Germany: Springer, 2015, pp. 354–368.

[111] H. Kwon, Y. Kim, H. Yoon, and D. Choi, "CAPTCHA image generation systems using generative adversarial networks," *IEICE Trans. Inf. Syst.*, vol. 101, no. 2, pp. 543–546, 2018, doi: 10.1587/transinf.2017EDL8175.

[112] Y. Matsuura, H. Kato, and I. Sasase, "Adversarial text-based CAPTCHA generation method utilizing spatial smoothing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[113] H. Kwon, H. Yoon, and K.-W. Park, "CAPTCHA image generation: Two-step style-transfer learning in deep neural networks," *Sensors*, vol. 20, no. 5, p. 1495, Mar. 2020.

[114] G. Ye, Z. Tang, D. Fang, Z. Zhu, Y. Feng, P. Xu, X. Chen, J. Han, and Z. Wang, "Using generative adversarial networks to break and protect text CAPTCHAs," *ACM Trans. Privacy Secur.*, vol. 23, no. 2, pp. 1–29, 2020.

[115] R. Shao, Z. Shi, J. Yi, P. Chen, and C. Hsieh, "Robust text CAPTCHAs using adversarial examples," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 1495–1504, doi: 10.1109/BigData55660.2022.10021100.

[116] M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, and D. Pérez-Cabo, "No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 11, pp. 2640–2653, Nov. 2017.

[117] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, M. F. Valstar, A. P. French, and T. P. Pridmore, Eds. Nottingham, U.K.: BMVA Press, 2014. [Online]. Available: http://www.bmva.org/bmvc/2014/papers/paper054/index.html

[118] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[119] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 818–833.

[120] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Y. Bengio and Y. LeCun, Eds., 2014, pp. 1–6.

[121] C. Shi, X. Xu, S. Ji, K. Bu, J. Chen, R. Beyah, and T. Wang, "Adversarial CAPTCHAs," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6095–6108, Jul. 2022.

[122] D. Hitaj, B. Hitaj, S. Jajodia, and L. V. Mancini, "Capture the bot: Using adversarial examples to improve CAPTCHA robustness to bot attacks," *IEEE Intell. Syst.*, vol. 36, no. 5, pp. 104–112, Sep. 2021.

[123] K. O. Stanley, "Compositional pattern producing networks: A novel abstraction of development," *Genetic Program. Evolvable Mach.*, vol. 8, pp. 131–162, May 2007.

[124] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665*.

[125] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–14.

[126] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[127] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[128] P. Tian, W. Liao, T. Kimbrough, E. Blasch, and W. Yu, "Generating adversarial robust defensive CAPTCHA (GARD-CAPTCHA) in convolutional neural networks," in *Proc. Int. Conf. Softw. Eng. Res. Appl.* Cham, Switzerland: Springer, 2022, pp. 17–31.

[129] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*.

[130] X. Jia, J. Xiao, and C. Wu, "TICS: Text–image-based semantic CAPTCHA synthesis via multi-condition adversarial learning," *Vis. Comput.*, vol. 38, no. 3, pp. 963–975, 2022.

[131] H. Park, Y. Yoo, and N. Kwak, "MC-GAN: Multi-conditional generative adversarial network for image synthesis," 2018, *arXiv:1805.01123*.

[132] J. Y. Shim, S. Jung, J. Kim, and J.-K. Kim, "Stabilized performance maximization for GAN-based real-time authentication image generation over internet," *Multimedia Tools Appl.*, vol. 83, no. 22, pp. 62045–62059, Jul. 2023.

[133] N. Dinh, K. Tran-Trung, and V. Truong Hoang, "Augment CAPTCHA security using adversarial examples with neural style transfer," *IEEE Access*, vol. 11, pp. 83553–83561, 2023.

[134] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," 2017, *arXiv:1705.06830*.

[135] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[136] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2921–2926.

[137] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[138] N. D. Trong, T. H. Huong, and V. T. Hoang, "New cognitive deep-learning CAPTCHA," *Sensors*, vol. 23, no. 4, p. 2338, Feb. 2023.

[139] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[140] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, Stockholm, Sweden, J. Lang, Ed., 2018, pp. 4208–4215, doi: 10.24963/ijcai.2018/585.

[141] M. Behjati, S.-M. Moosavi-Dezfooli, M. S. Baghshah, and P. Frossard, "Universal adversarial attacks on text classifiers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7345–7349.

[142] X. Dong, J. Han, D. Chen, J. Liu, H. Bian, Z. Ma, H. Li, X. Wang, W. Zhang, and N. Yu, "Robust superpixel-guided attentional adversarial attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12895–12904.

[143] Y. Wang, S. Wu, W. Jiang, S. Hao, Y. Tan, and Q. Zhang, "Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity," in *Proc. 13th Int. Joint Conf. Artif. Intell. (IJCAI)*, Montreal, QC, Canada, Z. Zhou, Ed., 2021, pp. 3125–3133, doi: 10.24963/ijcai.2021/430.

[144] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, "Frequency-driven imperceptible adversarial attack on semantic similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15315–15324.

[145] Z. Wang, M. Song, S. Zheng, Z. Zhang, Y. Song, and Q. Wang, "Invisible adversarial attack against deep neural networks: An adaptive penalization approach," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 3, pp. 1474–1488, May 2021.

[146] Z. Zhao, Z. Liu, and M. Larson, "Adversarial image color transformations in explicit color filter space," 2020, *arXiv:2011.06690*.

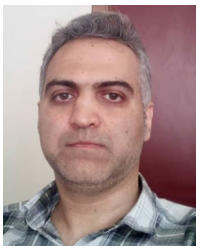
[147] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, X. Wang, and Y. Qiao, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 14408–14419.

[148] J. Fan, Y. Gao, H. Luo, and G. Xu, "Automatic image annotation by using concept-sensitive salient objects for image content representation," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, Eds. Sheffield, U.K.: ACM Press, 2004, pp. 361–368, doi: 10.1145/1008992.1009055.

[149] J. Wang, S. Zhu, J. Xu, and D. Cao, "The retrieval of the beautiful: Self-supervised salient object detection for beauty product retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2548–2552.

[150] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning," 2017, *arXiv:1703.01290*.
[151] Z. Yu, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7223–7233.
[152] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10076–10085.
[153] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4654–4662.
[154] T. Chen and J. Luo, "Expressing objects just like words: Recurrent visual embedding for image-text matching," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10583–10590.
[155] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10941–10950.
[156] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1218–1226.
[157] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang, "FashionBERT: Text and image matching with adaptive loss for cross-modal retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 2251–2260.
[158] N. Aziere and S. Todorovic, "Ensemble deep manifold similarity learning using hard proxies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7299–7307.
[159] J. Lu, C.-X. Ma, Y.-R. Zhou, M.-X. Luo, and K.-B. Zhang, "Multi-feature fusion for enhancing image similarity learning," *IEEE Access*, vol. 7, pp. 167547–167556, 2019.
[160] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 23–79, Aug. 2021.
[161] Z. Wei, P. Paliyawan, and R. Thawonmas, "Improving deep-feature image similarity calculation: A case study on an Ukiyo-e card matching game lottery," *IEEE Access*, vol. 10, pp. 44608–44616, 2022.
[162] W. Ma, Y. Li, X. Jia, and W. Xu, "Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4630–4639.
[163] B. Tian, F. Juefei-Xu, Q. Guo, X. Xie, X. Li, and Y. Liu, "AVA: Adversarial vignetting attack against visual recognition," 2021, *arXiv:2105.05558*.
[164] C. Hu, W. Shi, and L. Tian, "Adversarial color projection: A projector-based physical-world attack to DNNs," *Image Vis. Comput.*, vol. 140, Dec. 2023, Art. no. 104861.

MORTEZA MORADI received the B.S. degree in electrical engineering (control engineering) from the Qom University of Technology and the M.S. degree in electrical engineering (control engineering) from the University of Zanjan. He is currently pursuing the Ph.D. degree in systems, energy, computer and telecommunications engineering with the DIEEI-PeRCeiVe Laboratory, University of Catania. He was formerly with the Young Researchers and Elite Club, Qazvin Branch, and the SYNTECH Technology and Innovation Center. His research interests include deep learning, computer vision, and intelligent vehicles.



MOHAMMAD MORADI received the B.S. degree in computer engineering (software engineering) from the Ghazali Higher Education Institute and the M.S. degree in computer engineering (software engineering) from QIAU. He is currently pursuing the Ph.D. degree in systems, energy, computer and telecommunications engineering with DEEI, University of Catania. He was a Researcher with the Young Researchers and Elite Club, Qazvin Branch, and the SYNTECH Technology and Innovation Center. His major research interests include deep learning, human–computer interaction, and crowdsourcing.

SIMONE PALAZZO received the Ph.D. degree from the University of Catania, Italy, in 2017, with a thesis on human–machine interaction modalities for object segmentation and categorization in images and videos. He is currently an Assistant Professor with the University of Catania. His current research interests include medical image analysis, continual learning, video object segmentation, and scene understanding.

FRANCESCO RUNDO received the degree in computer science engineering and the Ph.D. degree in applied mathematics for technology from the University of Catania. He is currently a Senior Technical Staff Team Leader with STMicroelectronics, Catania. He is also a member of the Automotive Research and Development Power and Discretion Division, STMicroelectronics. He is also the Team Leader and the Project Leader regarding the development of artificial intelligence-based solutions (hardware and software) for automotive, industrial, and medical applications. He is a member of the Computer Science Ph.D. Scientific Board, Department of Mathematics and Computer Science, University of Catania. He is also a member of the Computer Science Ph.D. Scientific Board, National Ph.D. Program of Artificial Intelligence. He has co-authored more than 100 contributions in international journals, conference proceedings contributions, SI series, posters, abstracts, and lectures. He is also the co-inventor of several international patents. His main research interests include advanced bio-inspired models, advanced and perceptual deep learning, embedded systems for deep learning algorithms, advanced deep learning, and mathematical modeling for automotive, industrial, and healthcare applications. He is a member of several international conference program committees. He serves as a reviewer and the guest editor for several special issues organized by such key editors in the field of computer science. He serves as an Associate Editor for *IET Networks* and *Applied Computational Intelligence and Soft Computing* (Hindawi), a Research-Topic Editor for *Frontiers in Computer Science* and *Frontiers in Neuroinformatics*, and a Topic Editor for *Electronics* and *Drones*.

CONCETTO SPAMPINATO is currently an Associate Professor with the University of Catania, Italy. He is also a Courtesy Faculty Member with the Center for Research in Computer Vision, University of Central Florida. In 2014, he created and currently leads the Pattern Recognition and Computer Vision Laboratory (PeRCeiVe Lab). His research interests include machine learning and its application in multiple domains from medical image analysis to autonomous robot navigation. He is an Associate Editor of *Computer Vision and Image Understanding*, IEEE TRANSACTIONS ON MULTIMEDIA, and *Machine Vision and Applications*, as well as an Area Chair for multiple top-tier conferences, including CVPR 2022.

Open Access funding provided by 'Università degli Studi di Catania' within the CRUI CARE Agreement