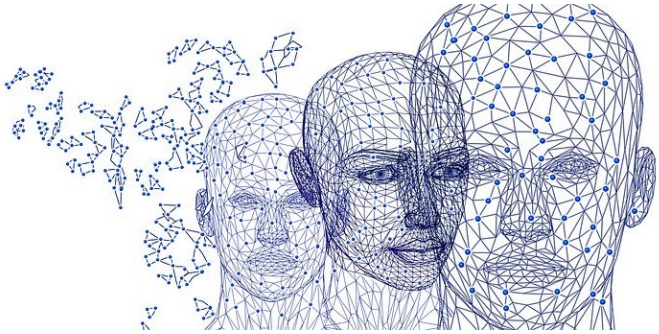
September 2020, Issue Code NL-2021-1

DDD (Digital Data Deception) Technology Watch Newsletter

Table of Contents



"*All warfare is based on deception. Hence, when we are able to attack, we must seem unable; when using our forces, we must appear inactive; when we are near, we must make the enemy believe we are far away; when far away, we must make him believe we are near.*"

— Sun Tzu, *The Art of War*

Editors: Enes Altuncu, Virginia Franqueira, Sanjay Bhattacherjee, and Shujun Li
Affiliation: Kent Interdisciplinary Research Centre in Cyber Security (KirCCS) & School of Computing, University of Kent, UK
Contact Us: ddd-newsletter@kent.ac.uk

Editorial

Deception has been used since ancient times, for both offensive and defensive purposes. In today's digital, and highly networked, world many deception-related activities happen in the cyber space and involve the creation of deceptive data in digital format. The use of digital data deception (DDD), by adversaries, call for countermeasures which include methods for detection of such deceptive data, as well as defensive deception that can mislead adversaries. The DDD Technology Watch Newsletter project has been established to monitor recent research and innovation progresses in DDD.

In this first issue, we focus on AI-related DDD, covering both offensive and defensive aspects. In future issues, we plan to cover other areas of DDD, including psychology-related DDD, and DDD lying between psychology (human) and AI (machine). We will also consider focusing on a more narrowly defined topic in the future, e.g., to cover a new and fast evolving technology in the field.

This issue is based on 37 research papers published in 2019 and 2020, which were identified and selected following a systematic literature review (SLR) approach called PRISMA (`http://prisma-statement.org/`). The sources used for searching relevant research papers were *Scopus* and *Web of Science*. We also identified three additional papers, deemed relevant, from other sources. We considered mainly peer reviewed papers, but decided to include a recent survey which has only been published as a pre-print on arXiv.org. As it is common for SLRs, we have not conducted peer-review of the papers. The summaries included in this newsletter are based on our understanding of the papers, assuming they are technically correct.

The papers included in this newsletter are organised into several sections, based on a number of relevant topics for AI-related DDD: 1) surveys on AI-related DDD; 2) spoofing AI authentication systems; 3) adversarial AI; and 4) detection of photo-realistic (fake) images. The first topic covers three surveys that provide a general overview of the area and its current scientific coverage. For the second topic, we consider generation of fake inputs for spoofing (i.e., failing) AI models. A special class of AI models – biometrics-based authentication systems – has been the target of many spoofing attacks. Therefore, we have a dedicated sub-topic covering that. For the third topic, we consider new methods for creating adversarial AI, with two sub-topics: creation of adversarial samples (i.e., attacks) and countermeasures against adversarial samples. For the last topic, we look at countermeasures against photo-realistic fake images, often created using deepfake-related techniques. Our classification of papers is based on a simple typology and there are overlaps between different topics, particularly for biometrics-related papers (some of which could be classified into more than one section).

We hope you enjoy reading this issue. Feedback is always welcome, and should be directed to `ddd-newsletter@kent.ac.uk`.

List of Acronyms

Throughout the newsletter, we used some common acronyms repeatedly. They are listed next, in alphabetic order, to avoid duplication.

AGN: Adversarial Generative Net(work)
AI: Artificial intelligence
CNN: Convolutional Neural Network
DDD: Digital Data Deception
DL: Deep Learning
DNN: Deep Neural Network
FGSM: Fast Gradient Sign Method
GAN: Generative Adversarial Networks
IDS: Intrusion Detection System
ML: Machine Learning
NN: Neural Network
PCA: Principal Component Analysis
RL: Reinforcement Learning
SVM: Support Vector Machine

Three Recent Surveys

DDD can be found in many different application contexts, e.g., spoofing of different kinds of AI systems using fake or manipulated inputs, anti-forensics used by criminals to remove, hide and fabricate data to mislead digital forensic processes, mis- and disinformation that cause spread of false information among people, and the use of defensive deception to protect a system. While deceptive data can always be created manually, automated approaches are always preferred especially for large-scale deception efforts. This section presents three surveys covering three different aspects of DDD with AI relevance.

Survey #1: Game Theory in Defensive Deception

Since deception always involves at least two partners (the one who deceives and the one who is deceived), game theory is a relevant technique to simulate interactions and potential moves considering both offensive and defensive deception as either side can adapt their behaviours according to the other party's behaviour.

Pawlick et al. [26] surveyed research works published between 2008 and 2018 that use game theory to model defensive deception. They identified the following six types of deception in the cyberspace.

Perturbation: Defences in this category typically use noise to limit leakage of sensitive information, for example, the adoption of a "differential privacy" or a "distortion privacy" approaches. The former purposely leaks a subset of information without advertising the actual level of privacy achieved, while the latter relies on measures of the attacker's inferences based on the information leaked.

Moving target defence: Defences in this category rely on agility and randomness such as changing attack surfaces and random configurations for networks, assets, and defence tools, e.g., using the use of a probability distribution of the IDS's location to limit the effectiveness of attackers' reconnaissance techniques.

Obfuscation: Defences in this category rely on hiding valuable information. For example, directing attackers to decoy targets rather than real assets, or revealing useless information.

Mixing: Defences in this category aim to prevent linkability. For example, via the use of pseudonym swapping, anonymity enhanced technologies, and exchange systems, such as mix networks and mix zones.

Honey-x: This category covers defences such as honeypots, honey-users, and honeybots. They aim to mimic valuable assets to attract attackers who are then monitored and studied.

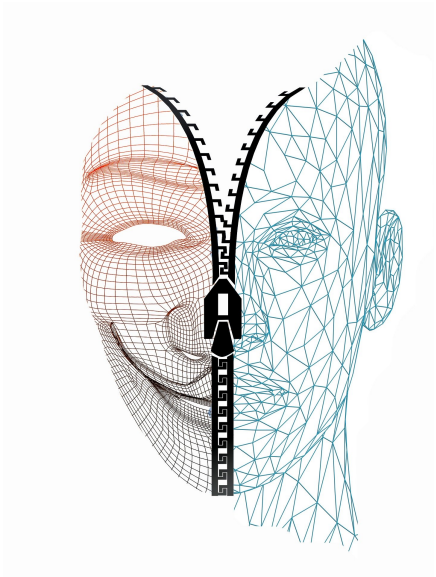
Attacker engagement: Defences in this category use feedback to dynamically influence attackers over an extended period of time. The aim is to waste their resources and gather intelligence.

Survey #2: Deepfake

Deepfake is a collective term used to describe a range of DL-based techniques to create fake media such as digital images,

videos and audio. This term is also used to refer to the fake media created by such methods. For example, a deepfake creation process may extract the feature set of a real face (i.e., of an existing individual) and apply those features to fabricate a fake face (i.e., of a non-existent individual). Ultimately, deepfake methods can create fake media, such as realistic-looking videos, by composing and changing objects, actions and voice. The proliferation of easy-to-use applications for deepfake creation, and the high quality of their output, shifted production of deepfake from the professional sphere to end-users.

Nguyen et al. [24], in a recent arXiv.org pre-print (2nd version published in July 2020), surveyed advances on deepfake creation and detection. The survey is comprehensive (149 papers cited), but not systematic because there is no mention of the methodology followed, such as how sources were identified and how papers were screened.



The authors classified deepfake detection into 2 classes: "fake image detection", and "fake video detection". The latter class is further sub-classified into "temporal features across video frames", including methods which utilise spatio-temporal features of video streams to detect deepfakes, and "visual artefacts within video frames", including methods which decompose videos into frames and explore visual artefacts within single frames to obtain discriminant features.

The authors' main conclusions in relation to deepfake detection were as follows.
(1) The quality of deepfakes has been increasing, and the performance of detection methods needs to improve accordingly.
(2) Detection methods are still in their early stages and the ones that have been proposed and evaluated are only using fragmented datasets.
(3) Current detection methods mostly focus on drawbacks of the deepfake generation pipelines, i.e., focus on finding weakness of the competitors to attack them. This kind of information and knowledge is not always available in adversarial environments.

The authors also raised the issue of authenticity verification of images and videos involved on an investigation. Experts opinion in a Court of Law by digital media forensics practitioners are no longer enough to determine accurately and reliably whether a digital media evidence is genuine or fake. Not even computer professionals can properly explain results of their deep learning methods, and this calls for research and advances in *Explainable AI*.

Survey #3: Textual Deception

Stylometry is a field of study that uses linguistic information in a text to extract non-linguistic properties such as identity, gender, age or occupation of the writer/author. The Human Stylome Hypothesis (HSH) conjectures that the author's identity can be reliably inferred from their stylistic "fingerprint" called *stylome*. *Adversarial stylometry* consists of techniques that attempt to defeat author identification or profiling. It aims to retain semantic content as much as possible to fool the classifier.

Gröndahl and Asokan [9] conducted a literature review of advances and trends in text deception (i.e., obfuscation techniques) and detection (i.e., author identification). They reviewed empirical works to show that certain linguistic features have been indicative of deception in certain collections of writings (or *corpora*). However, these features fail to generalise across divergent semantic domains. The authors suggest that deceptiveness does not leave content-invariant stylistic trace. Instead, textual similarity measures are better at classifying texts as potentially deceptive.

The paper also elaborates on forms of deception beyond semantic content. They focus on hiding author identity by *writing style obfuscation*. Through their survey on both author identification and

obfuscation techniques, they concluded that currently known style transformation methods fail to achieve reliable obfuscation while simultaneously ensuring that the transformed text is semantically close to the original text.

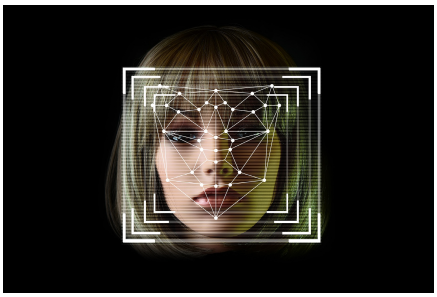
GAN-based approaches have been identified as promising for both adversarial stylometry, and automatic classification of text based on (say) "author gender" and "author age". The overall conclusions of the survey were:
(1) There is no evidence that deception leaves a content-invariant stylistic trace; so, to detect deception, semantic content should be compared across texts.
(2) Whether the HSH conjecture is true or not is uncertain; however, the deanonymisation attack is a realistic privacy concern.
(3) As of yet, automatic style transformation techniques do not ensure semantic faithfulness.

Spoofing Biometric Authentication Systems

User authentication via biometric features is very common for security applications. However, it is subject to *presentation attacks* which happen when a fake or forged biometric trait (e.g., fingerprint, iris, finger-vein or face) is displayed to a sensor in order to mislead the authentication system. A typical defence against presentation attack is *liveness detection*, used to determine if the presented biometric trait is an alive part of an individual or not.

In the following, we categorise papers falling into spoofing of biometric authentication systems, covering different biometric modalities.

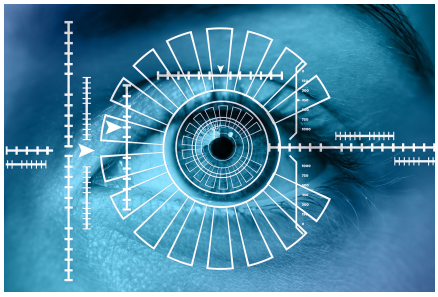
Face



Qu et al. [28] introduced a shallow CNN with Laplacian embedding (shallowCNN-LE) to detect the face liveness by making use of depth features and dynamic texture features. The authors compared the proposed method with several state-of-the-art approaches using CASIA, Replay-Attack and MSU USSA datasets, in terms of different measures.

Pinto et al. [27] applied the Shape-from-Shading (SfS) technique to detect face presentation attacks by recovering intrinsic properties of the scene, such as *albedo* (i.e., the proportion of light falling on a surface and irregularly reflected from it), depth, and reflectance properties of the facial surfaces. SfS is typically used to reconstruct the shape of an object from shading information in its surface. The properties they used were given to a CNN model as input for learning. The proposed method was evaluated with the Replay-Attack, CASIA and UVAD datasets, and compared with ResNet, SpoofNet and Xception architectures, and several proposed methods in terms of different measures.

Iris



Arora and Bhatia [1] proposed and evaluated a DL-based approach to deal with three different kinds of spoofing attacks: (1) printout of iris images, (2) textured contact lenses, and (3) photographs of paper printouts of images acquired for eyes wearing textured contact lenses. They preprocessed the iris images using the Hough transform to determine the centre coordinates and radius of the pupil and iris. Then, they used a DL architecture for feature extraction with a Softmax classifier to decide whether the iris is live or spoofed, and predict the type of attack. Experiments with the LivDet 2017 Iris competition dataset showed that with the introduction of more than one attack, the classification error in-

creases significantly. The authors concluded that the ability to detect more than one type of attack outperformed traditional systems which are trained to detect only a particular kind of attack.

Choudhary et al. [5] proposed what they called a Densely Connected Contact Lens Detection Network (DCLNet) aiming to reach a balance between ContlensNet and GHCLNet – state-of-the-art DL architectures for detection of presentation attacks via the use of contact lens. As such, they wanted to avoid iris normalisation and pre-processing and, at the same time, have a simpler network with comparable performance. After the feature extraction stage, they used a SVM classifier to discriminate between "no lens", "soft lens" and "coloured lens". DCLNet was evaluated using public datasets (ND Cosmetic Contact Lens 2013 and IIITD Contact Lens Iris) and different settings. The authors concluded that the DCLNet is a less complex model with optimal layer configuration, containing fewer learning parameters while exhibiting comparable performance with the state-of-the-art approaches for iris spoofing detection.

Fingerprint

Fei et al. [7] evaluated the robustness of several state-of-the-art fingerprint liveness detection models. They generated adversarial samples using white-box and black-box attacks in various settings, demonstrating the vulnerability of the models in those settings. For example, their evaluation has shown that the models tested (FGSM, MI-FGSM, and Deepfool) were not robust enough to transformations such as resize, horizontal flip, and rotation. They show that these schemes are likely to classify fake fingerprints as live fingerprints by adding tiny perturbations, even in a black-box setting. Based on these results, they proposed an enhanced adversarial attack algorithm to generate adversarial samples that are more robust to various transformations. The idea of the algorithm was to add a slight Gaussian noise in order to disturb the sample at every iteration, increasing the overall robustness. They used a number of public datasets, and achieved a higher success rate in their attacks, compared to other advanced methods.

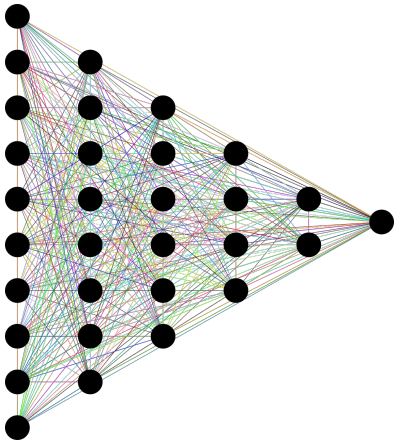
Sharma and Dey [30] proposed a static software-based approach using quality features to detect the liveness in a fingerprint image. The proposed method uses a 13-dimensional feature vector by extracting eight sensor-independent quality features from the detailed ridge–valley structure of a fingerprint at the local level. They tested their method on the publicly available database LivDet 2009. Experimental results demonstrated that the least average classification error of 5.3% was achieved on the database, indicating that the proposed method outperforms the current state-of-the-art approaches.

Finger-vein

Bok et al. [3] proposed a method to detect forged finger-vein using a technique called remote photoplethysmography (PPG). Finger-vein recognition uses the geometric information of blood vessels inside a finger' skin. A finger-vein reader equipped with a near infrared (NIR) camera was designed by the authors to build a database used in experiments. The dataset was composed of finger-vein video clips (579 real and 560 fake). The chance of a presentation attack in practice arises when images/videos captured by the NIR camera leak. By using PPG, the authors were able to collect time-series signals from the change of brightness in finger-vein video clips which correlates to heartbeat signals. This collection of features from fake and real finger-vein videos were fed into a Support Vector Machine (SVM) classifier. Results obtained were promising and achieved a 96.46% classification accuracy.

Spoofing Other AI Systems

The presentation attacks on biometric authentication systems can be applied to many other AI systems. The basic idea here is to present a fake or manipulated input that will mislead the target classifier. In this section, we cover papers about such spoofing attacks against other AI systems.

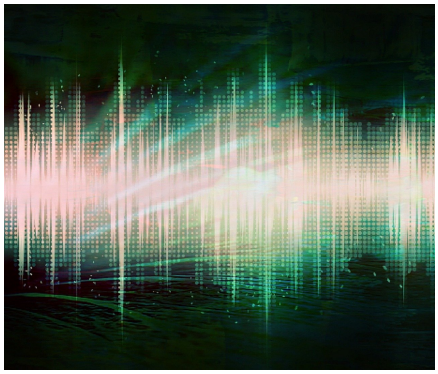


Since most of the works in this section use GANs, we start with a brief overview on the topic. GANs were introduced in 2014 by Goodfellow et al. [8]. A GAN is a framework for estimating generative models via an adversarial process. Two models are trained simultaneously. A generative model or *generator* G tries to capture the potential distribution of real samples, and generates new data samples. On the other hand, a discriminative model or the *discriminator* D works as a binary classifier estimating the probability that a sample came from the training data rather than being generated by G. The training procedure for G is to maximise the probability of D making a mistake. This interaction between G and D is nothing but a two-player zero-sum game between them. The gain or loss of one player is exactly balanced by the corresponding loss or gain of the other player. The optimisation goal is to reach the Nash equilibrium, where the generator is considered to have captured the distribution of real samples. For arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to $\frac{1}{2}$ everywhere (failing to distinguish between the samples).

Anti-Forensics

Anti-forensics are studies aiming to deceive forensic analysis. There are several mechanisms to bypass forensics, e.g., by tampering traces, by hiding traces, or erasing traces – ultimately avoiding detection.

Li et al. [16] proposed a GAN-based architecture to deceive state-of-the-art methods used for Audio Source Identification (ASI). ASI is relevant in the field of digital audio forensics. The overall goal of this work was to falsify forensic traces of recording devices in a way that the "attacked audio" can spoof the source identification classifier, while maintaining the appearance of authenticity and no perceptual traces of tampering. Two types of attacks were considered: *confusing attacks*, aiming to fabricate an audio very similar (according to a metric) to the original one, and *misleading attacks*, aiming to induce misclassification of a fabricated audio as belonging to a target category, chosen by the adversary. The GAN that was used incorporated a pre-trained ASI **detector** that feeds back to the generator helping it learn how to modify the forensic traces. For evaluation, the authors combined two public datasets (TIMIT-RD and LIVE-RECORD), and used three ASI detectors proposed in the literature. They reported a significant reduction of detection accuracy (from 97% to 5%) for confusing attacks, and a successful attack rate of 81.32% using six target categories for misleading attacks.



Wu et al. [34] proposed a GAN-based model to deceive state-of-the-art methods used to detect manipulated images. These techniques are especially relevant to the field of multimedia forensics. The authors claimed that existing image anti-forensic methods concentrate on concealing traces of a single operation (e.g., compression, blurring or splicing), rather than multiple operations, as it often happens in practice. Their model used a *generator (DL) network*, responsible for the generation of manipulated images resulting from multiple operations over the original images, and a *critic (GAN) network*, responsible for maximising the Wasserstein distance between the distributions of the manipulated images and the original ones. The model was evaluated over two public datasets (BossBase and BOWS2-Original) and two state-of-art forensic detectors, against 12 kinds

of two-manipulation chains and 8 kinds of image three-manipulation chains. They observed that the accuracy of the detectors deteriorated from around 99% (for one-manipulation) to around 50% (for two- or three-manipulation chains).

Wu et al. [33] proposed a Deep Reinforcement Learning based framework to deceive state-of-the-art botnet detection methods, relevant for the field of network forensics. The aim was to generate adversarial botnet flows (i.e., add perturbations to the flows and change their spatial and temporal properties) under black-box attack conditions, assuming a decision-based approach in place for detection. Their approach takes advantage of reinforcement learning (i.e., Deep Q-Network) consisting of an agent and an environment. The *agent* generates manipulated flows using Markov decision processes, and receives feedback (reward: benign or botnet, and state: feature vector of the sample) from the environment. The *environment* extracts features from the manipulated flows and uses an auto-encoder, i.e., a type of feed-forward neural network, to map flows to short binary codes and to learn efficient data encoding in an unsupervised manner. The authors evaluated their framework against two kinds of botnet detection models: a decision tree model based on 12 types of manually selected features, and a DL model based on CNN. Results with the "Malware Capture Facility Project" dataset showed that their botnet flow adversarial samples achieve detection evasion rates of 35-50% for the CNN detection model, while the average evasion rate for the decision tree detection model was less than 10%.

Spoofing Detection

Mohamed and Khalida [18] proposed a supervised machine learning-based approach to detect timestamp tampering in the NTFS filesystem. In order to generate a synthetic dataset for experiments, they created a Windows 10 virtual machine, populated it with files and used available online tools to tamper their timestamps using random time values, mimicking conditions of practice. To avoid bias, they applied a down-sampling strategy to re-balance the dataset, and randomly split the dataset using a 80/20 rule for training and testing. Features of interest were timestamp attributes extracted from the NTFS $MFT files of the dataset (Years, Months, Days number, Hours, Minutes, Seconds and Microseconds) converted to a numeric scale of [0,1]. They used a binary logistic regression ML algorithm to detect timestamp tampering. Evaluation metrics and confusion matrix indicated promising results, but they were not compared to the state-of-the-art; the authors briefly discussed limitations of the approach.

Fake MAV Navigation Data

Barbeau and Garcia-Alfaro [2] proposed a QGAN design to generate fake Micro Aerial Vehicle (MAV) navigation data, and discriminate fake and real MAV navigation data. They combined classical optimization, qubit quantum computing and photonic quantum computing. While Parrot Mambo MAV was used to generate genuine data, fake data was produced by using photonic devices in the photonic-quantum circuit. The performance of the proposed design was evaluated through a simulation on a classical computing platform.

Recent studies show that DL (especially supervised ML) models are vulnerable to adversarial attacks. Adversarial samples may be created by attackers to mislead a model at the time of training. A malicious sample or input $x' = x + \sigma$ is such that the slight perturbation σ (a very small quantity compared to x) makes no perceivable (visual or auditory) difference between x and x' for human beings. However, the models will be fooled and hence behave differently for x and x'. Such vulnerabilities are serious threats to their applications. Most attacks use an L_p-norm distance matrix to define the magnitude of the perturbation σ. For a perturbed image x' to be visually similar to x, each of the norms L_0, L_2 and L_∞ of σ have to be small. Given their importance, it is crucial to be able to generate adversarial samples not only for adversarial purposes, but also for improving the existing DL models. Several methods have been proposed for the creation of adversarial samples for different application domains.

In this section, we cover number of papers about creation methods of adversarial samples.

Embedding Attack

Liu et al. [17] presented an *embedding attack* in which adversaries can embed a small target image into a benign image to generate adversarial samples, that come into effect when the image is resized in the DL pre-processing pipeline. The embedding attack does not require any information about the network, and the only necessary information is the input image size. The effectiveness of the proposed attack was shown for three most common image resizing methods which are: nearest-neighbor, bilinear, and bicubic interpolation on Inception-v3 deep neural network.

A similar approach to the embedding attack was proposed by Chen et al. [4] under a different name (i.e., content disguising attack). They proposed three different attack approaches based on L_0, L_2 and L_∞ norm distance metrics. The authors tested the proposed method with different models, including Inception-v3, VGG-16, ResNet, Baidu Animal Classification, Aliyun Image Tagging and YOLO-v3 – which have different interpolation methods. The experimental results indicate that the proposed method gave 100% success rate in all settings.

Object Recognition

Sharif et al. [29] proposed a GAN-based general framework, called adversarial generative nets (AGNs), which can consider desired objectives when generating adversarial examples. These objectives could be beyond the similarity between adversarial input and the original image. They include robustness to ensure that the adversarial examples can fool the neural network in different imaging conditions (e.g., light and angle), inconspicuousness to avoid suspicion by human observers, and scalability to make a few examples sufficient to fool the neural network in several contexts. The proposed framework was demonstrated on two application domains: eyeglass frames – designed to deceive a face recognition system, and handwritten-digit classifier deception. While the target DL network to be deceived was trained on VGG and OpenFace datasets for the former domain, MNIST was used for the latter. Moreover, the authors constructed a dataset composed of 26,520 real images of eyeglasses by using Google search API to train the generator for the former domain. While they reported over 88-100% success rate for dodging and impersonation attacks against the face recognition system in digital and physical environments, the proposed framework managed to produce more than 5,000 adversarial digital images that appear comprehensible to human observers, but get misclassified by the DL network.

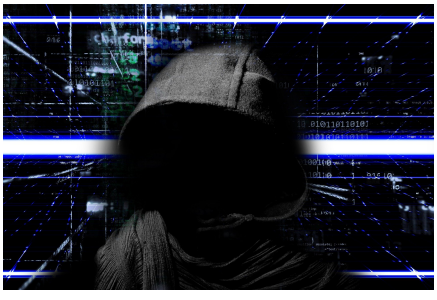
Zao et al. [40] presented systematic solutions to build robust and practical adversarial samples against real world object detectors. Particularly, for *Hiding Attack (HA)*, they proposed a Feature-Interference Reinforcement (FIR) method and Enhanced Realistic constraints Generation (ERG) to improve robustness. For *Appearing Attack (AA)*, they proposed the nested-AE, which combines two adversarial samples to

attack object detectors in both long and short distance. The authors also designed diverse styles of adversarial samples to make AA more surreptitious. Evaluation results show that the proposed adversarial samples can attack the state-of-the-art real-time object detectors (i.e., YOLO V3 and faster-RCNN) with a success rate of up to 92.4% with varying distance, from 1m to 25m.

Osahor et al. [25] applied FGSM against a VGG16-based DL target detector, which is capable of distinguishing target chips from background chips with 99% accuracy, to generate adversarial target images. The proposed method was evaluated with the Comanche FLIR dataset, which includes targets with different angular orientation and range. Experimental results indicated that the performance of the DL architecture dropped considerably by over 40% by generating adversarial target images with the proposed method.

Biometrics

In this subsection, we cover papers about generation of adversarial samples against biometric-related AI systems (but not focusing on user authentication applications, covered previously).



Kakizaki and Yoshida [12] proposed a method for generating unrestricted adversarial samples using image translation techniques. Their aim was to translate a source image into any desired facial appearance with large perturbations to deceive face recognition systems. More precisely, the authors considered hair colour, heavy makeup and eyeglasses for source image translation to generate adversarial samples. They observed an 80% attack success rate under black-box settings, and a 90% attack success rate under white-box settings in their experiments with the VGG and ResNet datasets. Moreover, to justify that the generated adversarial images were perceptually realistic, they showed 100 pairs of original and adversarial images to workers on Amazon Mechanical Turk, and asked if they believed they belonged to the same person or not; 76,6% of the workers answered that they showed the same person.

Xue et al. [35] proposed a face recognition neural network deceiving method based on the Deepfool algorithm [19] with the goal of having better privacy. In the proposed algorithm, adversarial samples generated to deceive face recognition systems are produced using a white-box attack and optimised with Euclidean distance to produce fake face images. The performance of the proposed algorithm was compared to the FSGM, which is a traditional image perturbation method. The comparison considered the Euclidean distance between the input image and the target category image on a dataset with 20 identities. Results showed that the proposed algorithm outperforms FSGM, although the processing time for each image was three times higher.

Kakizaki et al. [13] introduced a new method, namely GlassMasq, to produce adversarial examples with high confidence and small perturbation to deceive Deep Neural Network (DNN) based on face recognition systems using feature extractor. Their comparison between GlassMasq and the existing methods on two large datasets (more than three million images) show that it provides higher confidence and at most 62.8% smaller perturbation.

Nguyen et al. [23] proposed a method for generating face images to train presentation attack detection systems based on the CycleGAN network. Given the input face image, face detection was performed by using an Ensemble of Regression Tree (ERT) based method. This was followed by an in-plane rotation compensation performed as pre-processing steps. The authors evaluated the proposed method using the CASIA and Replay-mobile datasets, in terms of Frechet Inception Distance (FID) and presentation attack detection distance (padD), which is a measure they proposed.

Soleymani et al. [31] tried to mimic the iris-code generation filter bank procedure. This corresponds to the generation of the binarized iris template from an eye image, in iris recognition systems with a deep auto-encoder surrogate network to generate adversarial iris samples. They observed less than 2% error rate for the proposed surrogate network which according to the authors, means that it can mimic the conventional iris-code generation algorithms very closely. While the iris-code surrogate network was trained with BioCop dataset (composed of 10,000 pairs of normalised iris and mask images), the adversarial framework was tested with the BIOMDATA

dataset (composed of 3,040 iris images from 231 subjects). The authors evaluated the proposed method in terms of success rate with different step sizes, when maximum number of iterations changed.

Speech Recognition

Kwon et al. [14] proposed a selective audio adversarial sample with minimum distortion. The idea is to become misclassified (as the target phrase) by a victim classifier but correctly classified (as the original phrase) by a protected classifier. To generate such samples, a transformation is carried out to minimise the probability of incorrect classification by the protected classifier and that of correct classification by the victim classifier. The authors conducted experiments targeting the state-of-the-art DeepSpeech voice recognition model using Mozilla Common Voice datasets and the Tensorflow library. They showed that the proposed method can generate a selective audio adversarial example with a 91.67% attack success rate and a 85.67% protected classifier accuracy.

Text Classification

Zhang et al. [39] proposed a method that generates readable adversarial texts against text classification systems with some perturbations that can also confuse human observers successfully. The method utilised the Continuous Bag-of-Words (CBOW) model to train word embedding and look for appropriate perturbations to generate the adversarial texts through controlling the perturbation direction vectors. Once the adversarial texts were generated, they were gathered (with the original texts) to perform adversarial training for supervised learning or virtual adversarial training for semi-supervised learning (if real labels did not exist) using a Recurrent Neural Network text classifier. The authors tested the proposed method on the IMDB, Elec and Rotten Tomatoes datasets and it reached 93%, 94% and 83% accuracy rate, respectively.

Intrusion Detection Systems

Cordy et al. [6] proposed a search-based approach to test IDS by automatically generating training attacks. Going a step further, they proposed searching for countermeasures, learning from the successful attacks and, thereby, increasing the resilience of the tested IDS. The proposed approach was evaluated on a denial-of-service attack detection scenario and a dataset recording the network traffic of a real-world system. Experiments showed that the proposed search-based attack scheme generated successful attacks bypassing the current state-of-the-art defences. By co-evolving the proposed attack and defence mechanisms, the authors managed to improve the defence of the IDS under test by making it resilient to 49 out of 50 independently generated attacks.

In order to counter adversarial AI systems, a key step is the detection process. An increased accuracy of detection means the models become more robust. This has to be measured and compared with previously proposed models. Beyond that, there are other ways to train as well as evaluate the performance of the countermeasure techniques. In this section, we cover number of papers about defending adversarial AI.

Robustness Evaluation

Mundra et al. [21] proposed a pre-processing scheme for detection of adversarial images under four types of white-box attacks. Their approach involves two algorithms. The first one is a Principal Component Analysis (PCA) algorithm applied to the training dataset to extract vectors from RBG (red, blue, green) dimensions of the image pixels which were arranged from most significant to least significant components. The second algorithm is responsible for causing random perturbation on the least significant components of the image vectors, and provides a decision-based output (i.e., by determining if the input is adversarial or not). The authors evaluated the performance of their approach using the public dataset CIFAR-10, with different settings (e.g., number of perturbed image samples, and coefficient of perturbation). Their work has low computation complexity and low rate of false positives.

Yu et al. [36] proposed new defence mechanism against adversarial samples using the natural idea to output multiple results instead of one. Their method is based on appending information much like building a self-ensemble model. They proposed two algorithms - one basic and one score-based. They empirically showed that using their techniques, models can be made more robust against static white-box attacks compared to adversarial training models. In particular, they found that even in the case of a full white-box attack where an adversary can craft malicious examples from defence models, their method has a more robust performance of about 54.6% precision on the Cifar10 dataset and 38.7% precision on the Mini-Imagenet dataset. Another advantage of their method is that it is able to maintain the prediction accuracy of the classification model on clean images, and thereby exhibits its high potential in practical applications.

Wang and Qiao [32] proposed the large margin cosine estimation (LMCE) detection scheme to achieve robustness against adversarial samples. They modelled various types of adversarial attacks and established proposed defence mechanisms against them and evaluated their approach. They validated their method on a range of standard datasets including MNIST, CIFAR-10, and SVHN. The assessment strongly reflected the robustness of this approach in the face of various white and semi-white box attacks.

Detecting Adversarial Samples

Mun and Kang [20] proposed a random binary ensemble model that exploits multiple binary encoded labels to improve adversarial robustness of CNNs to white-box attack models. They used the random target encoding instead of traditional one-hot encoding method to represent the input class. Subsequently, the duplicates of the same CNN architectures were simultaneously trained with their own unique binary codes creating an ensemble and yet optimised through a single objective function. The distinct binary codes assigned for each input can result in different weights for each classifier. On the other hand, the classifiers also interact with others as they are trained with the same objective function. The random binary encoding method has multiple high bits compared to a simple one-hot code, and therefore an attacker can not readily predict the gradient. In summary, the randomness reduced the susceptibility by diversifying the directions into many dimensions.

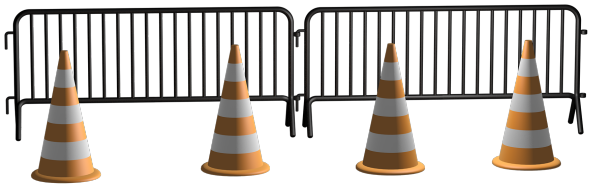
Mygdalis et al. [22] proposed a novel adversarial attack methodology for fooling DNN classi-

fiers for images and also provided a novel defence mechanism to counter such attacks. Two concepts were introduced: the K-Anonymity-inspired Adversarial Attack (K-A^3), and the Multiple Support Vector Data Description Defence (M-SVDD-D). Inspired by the K-Anonymity principles, the proposed concept $K - A^3$ introduced novel optimisation criteria to standard adversarial attack methodologies. The adversarial examples it generates, are not only misclassified by the neural network classifier, but are uniformly spread along K different ranked output positions. The proposed concept (M-SVDD-D) consists of a deep neural architecture layer and an additional class verification mechanism. This deep neural architecture layer is in turn made up of multiple non-linear one-class classifiers based on Support Vector Data Description that can be used to replace the final linear classification layer of a deep neural architecture. Its application increases the noise energy required to deceive the protected model and hence decreases the effectiveness of adversarial attacks. Here, the noise is nothing but the introduced non-linearity. In addition, M-SVDD-D can be used to prevent adversarial attacks in black-box attack settings.

Hashemi and Mozaffari [10] aimed to immunise DNNs through adversarial example generation and training so that evasion attacks are minimised. They proposed a GAN with a multi-class discriminator for producing a noise which when added to the original image, the adversarial examples can be obtained. In this paper, various types of evasion attacks have been considered and performances of the proposed methods are evaluated on different victim models under various defensive strategies. Experimental results were conducted based on MNIST and CIFAR10 datasets and the average success rates for different attacks have been reported and compared with state-of-the-art methods. The success rates of non-targeted attacks on DNNs after training by adversarial examples, reduced from 87.7% to 10.41% using the MNIST dataset and from 91.2% to 57.66% using the CIFAR-10 dataset.

The Use of Game Theory

Zhang and Zhu [37] proposed several defence strategies for a distributed support vector machine (DSVM) learner against a potential adversary. They capture the conflicting interests between the DSVM learner and the attacker through a nonzero-sum game to model their strategic interactions with a set of nodes. They use the Nash equilibrium of the game in predicting the outcome of learning algorithms in adversarial environments. They develop secure and resilient distributed algorithms based on alternating direction method of multipliers (ADMoM). They present four defence strategies against potential attackers - (1) to use balanced networks with fewer nodes and higher degrees, (2) adding training samples to compromised nodes (to reduce the vulnerability of the learning system) and to uncompromised nodes at the beginning of the training process to make the learner less vulnerable, and (3) to use verification methods where each node verifies its received data, and only accepts reasonable information from neighbouring nodes, and (4) use rejection method where each node rejects unacceptable updates. They show the effectiveness of these strategies using numerical experiments.



Detecting Photo-realistic (Fake) Images

Computer-generated photo-realistic images are widely used as a tool to deceive humans and spread mis- and dis-information. Therefore, a good way of detecting such deception is distinguishing them from those generated by a genuine imaging process (i.e., taken by a digital camera or obtained by scanning a picture on paper). In this section we cover three papers reporting detection methods of such images.

Detecting Computer-Generated Images

Zhang et al. [38] made use of channel and pixel correlation to distinguish computer-generated images from real images in their CNN-based model. They utilised the fact that camera sensors perform interpolation for each pixel to predict the missing colour information resulting from the filtering process by Colour Filter Array, which is not the case for computer-generated images. Based on this, they claimed that channel and pixel correlation caused by interpolation can be used to detect computer-generated images. While they designed a self-coding network to explore correlation between colour channels, they applied many convolutional layers without pooling operation to identify pixel correlation. The proposed network reaches around 94% classification accuracy in SPL2018 dataset and outperforms some state-of-the-art approaches such as LiNet, BSP-CNN and YaoNet.

He [11] proposed a method for classifying between real and computer generated images based on CNN through transfer learning. In this method, *transfer learning* was adopted to VGG and ResNet networks separately to boost the accuracy of both networks. The proposed model was evaluated with the DSTok dataset, and the results showed that transfer learning increases the accuracy rate of the VGG and ResNet networks from 71% and 75% to 92% and 96%, respectively.

Detecting Deep Network-Generated Images

Li et al. [15] addressed the problem of detecting deep network generated (DNG) images. They proposed a feature set based on the chrominance (difference between one colour and a reference colour of the same brightness and chromaticity) components in the residual domain to distinguish DNG images from real images. The study also considers different detection scenarios, including matched or mismatched training-test data and unknown generative model. The authors evaluated the proposed method with several datasets including face and bedroom images in both low and high resolution. The experimental results show that the proposed method reaches 99% average classification accuracy like the other compared methods. On the other hand, the proposed method outperforms other methods when training-test data are mismatched and the generative model is unknown.

References

[1] Shefali Arora and M. P. S. Bhatia. 2020. Presentation Attack Detection for Iris Recognition Using Deep Learning. *International Journal of System Assurance Engineering and Management* 11, 2s (2020), 232–238. https://doi.org/10.1007/s13198-020-00948-1

[2] Michel Barbeau and Joaquin Garcia-Alfaro. 2019. Faking and Discriminating the Navigation Data of a Micro Aerial Vehicle Using Quantum Generative Adversarial Networks. In *Proceedings of 2019 IEEE GLOBECOM Workshops*. IEEE. https://doi.org/10.1109/GCWkshps45667.2019.9024550

[3] Jin Yeong Bok, Kun Ha Suh, and Eui Chul Lee. 2019. Detecting Fake Finger-Vein Data Using Remote Photoplethysmography. *Electronics* 8, 9, Article 1016 (2019), 10 pages. https://doi.org/10.3390/electronics8091016

[4] Yufei Chen, Chao Shen, Cong Wang, Qixue Xiao, Kang Li, and Yu Chen. 2020. Scaling Camouflage: Content Disguising Attack Against Computer Vision Applications. *IEEE Transactions on Dependable and Secure Computing* (2020). https://doi.org/10.1109/TDSC.2020.2971601 in press.

[5] Meenakshi Choudhary, Vivek Tiwari, and Venkanna U. 2019. An Approach for Iris Contact Lens Detection and Classification Using Ensemble of Customized DenseNet and SVM. *Future Generation of Computer Systems* 101 (2019), 1259–1270. https://doi.org/10.1016/j.future.2019.07.003

[6] Maxime Cordy, Steve Muller, Mike Papadakis, and Yves Le Traon. 2019. Search-Based Test and Improvement of Machine-Learning-Based Anomaly Detection Systems. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ACM, 158–168. https://doi.org/10.1145/3293882.3330580

[7] Jianwei Fei, Zhihua Xia, Peipeng Yu, and Fengjun Xiao. 2020. Adversarial Attacks on Fingerprint Liveness Detection. *EURASIP Journal of Image and Video Processing* 2020, 1, Article 1 (2020), 11 pages. https://doi.org/10.1186/s13640-020-0490-z

[8] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[9] Tommi Gröndahl and N. Asokan. 2019. Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? *Comput. Surveys* 52, 3, Article 45 (2019), 36 pages. https://doi.org/10.1145/3310331

[10] Atiye Sadat Hashemi and Saeed Mozaffari. 2019. Secure Deep Neural Networks Using Adversarial Image Generation and Training with Noise-GAN. *Computers & Security* 86 (2019), 372–387. https://doi.org/10.1016/j.cose.2019.06.012

[11] Ming He. 2019. Distinguish Computer Generated and Digital Images: A CNN Solution. *Concurrency and Computation: Practice and Experience* 31, 12, Article e4788 (2019), 10 pages. https://doi.org/10.1002/cpe.4788

[12] Kazuya Kakizaki and Kosuke Yoshida. 2020. Adversarial Image Translation: Unrestricted Adversarial Examples in Face Recognition Systems. In *Proceedings of 2020 AAAI Workshop on Artificial Intelligence Safety*. http://ceur-ws.org/Vol-2560/paper4.pdf Also available as an online pre-print arXiv:1905.03421 [cs.CV] at https://arxiv.org/abs/1905.03421.

[13] Kazuya Kakizaki, Kosuke Yoshida, and Tsubasa Takahashi. 2019. GlassMasq: Adversarial Examples Masquerading in Face Identification Systems with Feature Extractor. In *Proceedings of 2019 17th International Conference on Privacy, Security and Trust*. IEEE, 258–264. https://doi.org/10.1109/PST47121.2019.8949019

[14] Hyun Kwon, Yongchul Kim, Hyunsoo Yoon, and Daeseon Choi. 2019. Selective Audio Adversarial Example in Evasion Attack on Speech Recognition System. *IEEE Transactions on Information Forensics and Security* 15 (2019), 526–538. https://doi.org/10.1109/TIFS.2019.2925452

[15] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. 2020. Identification of Deep Network Generated Images Using Disparities in Color Components. *Signal Processing* 174, Article 107616 (2020), 12 pages. https://doi.org/10.1016/j.sigpro.2020.107616

[16] Xiaowen Li, Diqun Yan, Li Dong, and Rangding Wang. 2019. Anti-Forensics of Audio Source Identification Using Generative Adversarial Network. *IEEE Access* 7 (2019), 184332–184339. https://doi.org/10.1109/ACCESS.2019.2960097

[17] Yujia Liu, Weiming Zhang, and Nenghai Yu. 2019. Query-free Embedding Attack Against Deep Learning. In *Proceedings of 2019 IEEE International Conference on Multimedia and Expo*. IEEE, 380–386. https://doi.org/10.1109/ICME.2019.00073

[18] Alji Mohamed and Chougdali Khalida. 2019. Detection of Timestamps Tampering in NTFS Using Machine. *Procedia Computer Science* 160 (2019), 778–784. https://doi.org/10.1016/j.procs.2019.11.011 Proceedings of the International Workshop on Emerging Networks and Communications (IWENC 2019).

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2574–2582. https://doi.org/10.1109/CVPR.2016.282

[20] Ye-Ji Mun and Je-Won Kang. 2019. Ensemble of Random Binary Output Encoding for Adversarial Robustness. *IEEE Access* 7 (2019), 124632–124640. https://doi.org/10.1109/ACCESS.2019.2937604

[21] Kartik Mundra, Arpan Chattopadhyay, and Indra Narayan Kar. 2020. Adversarial Image Detection in Cyber-Physical Systems. In *Proceedings of the First ACM Workshop on Autonomous and Intelligent Mobile Systems*. ACM, Article 3, 5 pages. https://doi.org/10.1145/3377283.3377285

[22] Vasileios Mygdalis, Anastasios Tefas, and Ioannis Pitas. 2020. K-Anonymity Inspired Adversarial Attack and Multiple One-class Classification Defense. *Neural Networks* 124 (2020), 296–307. https://doi.org/10.1016/j.neunet.2020.01.015

[23] Dat Tien Nguyen, Tuyen Danh Pham, Ganbayar Batchuluun, Kyoung Jun Noh, and Kang Ryoung Park. 2020. Presentation Attack Face Image Generation Based on a Deep Generative Adversarial Network. *Sensors* 20, 7, Article 1810 (2020), 24 pages. https://doi.org/10.3390/s20071810

[24] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2020. Deep Learning for Deepfakes Creation and Detection: A Survey. Online pre-print, arXiv:1909.11573v2 [cs.CV]. https://arxiv.org/abs/1909.11573

[25] Uche M. Osahor and Nasser M. Nasrabadi. 2019. Deep Adversarial Attack on Target Detection Systems. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, Vol. 11006. SPIE, Article 110061Q, 9 pages. https://doi.org/10.1117/12.2518970

[26] Jeffrey Pawlick, Edward Colbert, and Quanyan Zhu. 2019. A Game-theoretic Taxonomy and Survey of Defensive Deception for Cybersecurity and Privacy. *Comput. Surveys* 52, 4 (2019), 82–110. https://doi.org/10.1145/3337772

[27] Allan Pinto, Siome Goldenstein, Alexandre Ferreira, Tiago Carvalho, Helio Pedrini, and Anderson Rocha. 2020. Leveraging Shape, Reflectance and Albedo From Shading for Face Presentation Attack Detection. *IEEE Transactions on Information Forensics and Security* 15 (2020), 3347–3358. https://doi.org/10.1109/TIFS.2020.2988168

[28] Xiaofeng Qu, Jiwen Dong, and Sijie Niu. 2019. shallowCNN-LE: A Shallow CNN with Laplacian Embedding for Face Anti-spoofing. In *Proceedings of 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 145–152. https://doi.org/10.1109/FG.2019.8756569

[29] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2019. A General Framework for Adversarial Examples with Objectives. *ACM Transactions on Privacy and Security* 22, 3, Article 16 (2019), 30 pages. https://doi.org/10.1145/3317611

[30] Ram Prakash Sharma and Somnath Dey. 2019. Fingerprint Liveness Detection Using Local Quality Features. *The Visual Computer* 35, 10 (2019), 1393–1410. https://doi.org/10.1007/s00371-018-01618-x

[31] Sobhan Soleymani, Ali Dabouei, Jeremy Dawson, and Nasser M. Nasrabadi. 2019. Adversarial Examples to Fool Iris Recognition Systems. In *Proceedings of 2019 International Conference on Biometrics*. IEEE. https://doi.org/10.1109/ICB45273.2019.8987389

[32] Shen Wang and Zhuobiao Qiao. 2019. Robust Pervasive Detection for Adversarial Samples of Artificial Intelligence in IoT Environments. *IEEE Access* 7 (2019), 88693–88704. https://doi.org/10.1109/ACCESS.2019.2919695

[33] Di Wu, Binxing Fang, Junnan Wang, Qixu Liu, and Xiang Cui. 2019. Evading Machine Learning Botnet Detection Models via Deep Reinforcement Learning. In *Proceedings of the 2019 IEEE International Conference on Communications*. IEEE. https://doi.org/10.1109/ICC.2019.8761337

[34] Jianyuan Wu, Zheng Wang, Hui Zeng, and Xiangui Kang. 2019. Multiple-Operation Image Anti-Forensics with WGAN-GP Framework. In *Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1303–1307. https://doi.org/10.1109/APSIPAASC47483.2019.9023173

[35] Jingsong Xue, Yu Yang, and Dongsheng Jing. 2019. Deceiving Face Recognition Neural Network with Samples Generated by Deepfool. *Journal of Physics: Conference Series* 1302, 2, Article 022059 (2019), 9 pages. https://doi.org/10.1088/1742-6596/1302/2/022059

[36] Yueyao Yu, Pengfei Yu, and Wenye Li. 2019. AuxBlocks: Defense Adversarial Examples via Auxiliary Blocks. In *Proceedings of 2019 International Joint Conference on Neural Networks*. IEEE, Article N-2040, 8 pages. https://doi.org/10.1109/IJCNN.2019.8851823

[37] Rui Zhang and Quanyan Zhu. 2019. Game-Theoretic Defense of Adversarial Distributed Support Vector Machines. *Journal of Advances in Information Fusion* 14, 1 (2019), 3–21. http://confcats_isif.s3.amazonaws.com/web-files/journals/entries/01-JAIF_Vol14_1_3-40_0.pdf

[38] Rui-Song Zhang, Wei-Ze Quan, Lu-Bin Fan, Li-Ming Hu, and Dong-Ming Yan. 2020. Distinguishing Computer-Generated Images from Natural Images Using Channel and Pixel Correlation. *Journal of Computer Science and Technology* 35 (2020), 592–602. https://doi.org/10.1007/s11390-020-0216-9

[39] Wei Zhang, Qian Chen, and Yunfang Chen. 2020. Deep Learning Based Robust Text Classification Method via Virtual Adversarial Training. *IEEE Access* 8 (2020), 61174–61182. https://doi.org/10.1109/ACCESS.2020.2981616

[40] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. 2019. Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1989–2004. https://doi.org/10.1145/3319535.3354259