Physical Domain Adversarial Attacks Against Source Printer Image Attribution

Nischay Purnekar, Benedetta Tondi, Mauro Barni
University of Siena, Italy
E-mail: nischay.purnekar@student.unisi.it, benedetta.tondi@unisi.it, mauro.barni@unisi.it

***Abstract*—Deep learning networks are vulnerable to adversarial examples, specifically subtle, human-imperceptible modifications that can deceive them. While most research has focused on digital adversarial attacks, in several applications it is necessary that the adversarial examples operate in the physical domain. Physical domain adversarial examples are usually crafted to ensure that the artifacts introduced by the attack survive the digital-to-analog and analog-to-digital transformations involved in such attacks. In this paper, we introduce an approach to generate adversarial examples against a source printer attribution system, aiming to determine which printer was used to print a given image document. With respect to conventional physical domain attacks, attacking a source printer attribution system poses additional challenges since the subtle features the attribution network relies on are introduced again during the Print and Scan (P&S) process that follows the attack, thus possibly nullifying the attack. We address this challenge by applying Expectation Over Transformation, including within the pool of transformations a simulation of the P&S process relying on two Generative Adversarial Network models trained for this purpose. Experimental results demonstrate that our approach yields a significant increase in attack success rates, surpassing those of baseline models.**

I. Introduction

Despite their effectiveness, Artificial Intelligence (AI) systems based on Deep Learning (DL) are vulnerable to various malicious attacks, including adversarial examples [1], backdoor attacks [2], and inversion attacks [3]. Adversarial examples involve subtle, human-imperceptible perturbations that lead to misclassifications or other incorrect behaviors. Most research has focused on pixel-level digital adversarial examples [4], assuming the attacker has full control over the image's digital representation. In contrast, physical adversarial examples [5], [6] exploit variations in texture, shape, and lighting, processed through the system's sensor inputs. Examples include specific patterns applied to physical objects like stop signs, which cause misidentification by autonomous vehicle perception systems. Despite the potential risks, there is significantly less research on generating and defending against physical adversarial attacks compared to digital ones.

This paper focuses on attacks against image-printed document authentication, which is crucial for legal, governmental, and financial sectors that handle sensitive and confidential information. Ensuring document integrity is vital to prevent forgery and fraud, as these can have significant consequences. The Federal Trade Commission reported 2.6 million fraud cases, resulting in $10.3 billion in losses in 2023 [7] due to piracy. Maintaining printed document authenticity and confidentiality is essential for protecting sensitive information and upholding trust in official processes. Within this framework, the goal of this paper is to study the vulnerability of an image printer source attribution classifier based on DL against physical adversarial examples. The classifier is trained to identify a document's originating printer using a diverse set of documents from various printers. We aim to generate adversarial examples that remain effective after reprinting by applying different attack algorithms. Traditionally, adversarial examples in the physical domain are created by adding perturbations directly to digital images, which are then transformed into a physical document or 3D object and fed to the AI model, successfully misleading the system. In our case, the attacked digital images are printed again by the same printer and scanned back before being fed to the classifier. The Print and Scan (P&S) process applied to the attacked images poses several challenges to the creation of effective attacks. Firstly, the P&S process degrades the perturbation introduced by the attack, thus requiring a stronger perturbation. Secondly, and most importantly, the features the attribution network relies on are re-introduced when the attacked digital image is printed for the second time, possibly nullifying the effectiveness of the attack.

Following previous work on the generation of physical adversarial examples, we use Expectation Over Transformation (EOT) [8] to craft perturbations that survive the distortion introduced when transitioning from the physical to the digital domain. As shown in Sect. V, classical EOT alone is not sufficient to maintain the effectiveness of adversarial examples after the P&S process due to the reintroduction of printing artifacts on top of the attacked image. For this reason, we propose incorporating a P&S simulator within the EOT framework to generate an adversarial attack that preemptively accounts for the subsequent reprinting process. In particular, we used a Pix2Pix Generative Adversarial Network (GAN) [9] and a CycleGAN [10] to simulate the P&S transformation. We integrated EOT with P&S into the Iterative Fast Gradient Sign Method (IFGSM) and the Carlini & Wagner (C&W) attacks, achieving a high Attack Success Rate (ASR) even after reprinting.

Given the above, the main contributions of this work are:

1) We developed two P&S simulators utilizing Pix2Pix GAN and CycleGAN image translation models. The simulators are not only applicable for crafting adver-

sarial source printer image attribution but can also be potentially used in digital image forensics to enhance the robustness of synthetic image detectors [11].
2) We integrated the P&S simulators as an additional transformation step in the EOT attack.
3) We were able to generate robust adversarial examples that withstand the reprinting process, successfully deceiving the target source printer attribution classifier.

The paper is organized as follows: Sect. II reviews adversarial attacks in digital and physical domains. Sect. III details the development and performance of the P&S simulators. Sect. IV focuses on the generation of robust adversarial examples. Sect. V analyzes the experimental results. Sect. VI summarizes our findings and suggests directions for future work.

II. Related Work

Adversarial examples, first identified by Szegedy et al. [1], demonstrate that minor perturbations can significantly alter a network's output while remaining nearly imperceptible to humans. These examples often generalize across different models, even those trained with varying hyperparameters or architectures, sparking significant interest in DNNs. Current research explores adversarial examples in both digital and physical domains.

Digital Domain adversarial perturbations are directly applied to the network's input, constraining the l_p-norm (e.g., l_∞-norm [4], l_2-norm, and l_0-norm [12]) of the perturbation to be lower than a certain threshold to maintain the imperceptibility of the attack. Depending on the adversary's knowledge, the adversarial attacks can be categorized as either white-box or black-box. In white-box attacks, the attacker has complete knowledge of the model, allowing full use of the gradient to craft the perturbations [4], [5], [12]. In a black-box attack, the attacker can only query the target model and receive corresponding outputs without access to its internal structure. In this scenario, the attacker can either leverage the generalizability of adversarial examples across different models or deduce the model's internal information through multiple queries.

Physical Domain attacks were pioneered by Kurakin et al. [5], who introduced the first physical-domain attack by printing digitally perturbed images, which were then photographed with a smartphone and fed into a pre-trained Inception v3 classifier. Their results indicated a decline in the effectiveness of the attack after the images underwent printing and photography. Sharif et al. [6] created adversarial eyeglass frames to deceive facial recognition systems by incorporating a non-printability score (NPS) and total variation (TV) loss in their optimization, ensuring printer accuracy and smooth color transitions. [13] used a similar TV loss to generate adversarial stickers for hats to fool the ArcFace system. Lu et al. [14] noted that attack effectiveness diminishes when images are viewed from different angles and distances. To generate physical adversarial examples that withstand the transformations involved when going back and forth from the digital to the physical domain, EOT was introduced in [8]. Eykholt et al. [15] refined EOT with Robust Physical Perturbation (RP2), sampling synthetic and physical transformations to create adversarial stop signs using posters or stickers, although this requires printing and photographing the original image multiple times. Jan et al. [16] proposed D2P, a transformation using a conditional GAN [9], [10] before EOT to simulate printing and photographing effects, but it faces feasibility issues due to the need for extensive printing and photographing to build a training dataset. A work that is somewhat similar to the present work is [17]. Even there, the detector relies on the features that are reintroduced after rebroadcast hence requiring the design of a particular EOT strategy. However, the rebroadcasting artifacts are different from those introduced by P&S, hence the method proposed in [17] cannot be applied in our case.

As a matter of fact, all the attacks based on EOT include natural geometric and color transformations to generate robust adversarial examples. As we will show later, however, this is not enough when the target system is a printer source attribution model. For this reason, we integrated the P&S simulators into the EOT framework. In this way, we were able to significantly improve the ASR, ensuring that the attack remains effective even after reprinting.

III. Print and Scan Simulation

Printing and scanning an image involves converting the digital images to physical copies and back to the digital domain, introducing various distortions and artifacts. Printing can cause color shifts, ink diffusion, and minor geometric distortions due to the printer's mechanical characteristics and type of paper used. Scanning adds further distortions and noise depending on the scanner's resolution, color response, and mechanical misalignments. These steps affect pixel values and introduce artifacts specific to the printer and scanner, along with minor geometric alterations due to imperfect paper positioning within the scanner.

Given the time-consuming and costly nature of manually creating large volumes of printed and scanned images, we developed two P&S simulators to be directly included within the EOT process, enabling the vast generation of training images without the expense and effort of physical P&S. Research on simulating the P&S process by means of deep learning is sparse. A significant contribution in this domain comes from Ferrara et al. [18], who demonstrated that integrating a simulated P&S transformation during training improves the accuracy of face morphing attacks on printed and scanned face images. Their model estimates the pixel distortions incurred during printing and scanning, considering various critical parameters such as the responsivity of the acquisition device, the sampling function characterizing the digitization process of printed images, the point spread function of the printer and scanner, noise levels, and color transformations. However, the presence of device-dependent unknown parameters complicates real-world adaptations, as calculating the point spread functions of printers and scanners is challenging, and fine-tuning each parameter can be time-consuming, especially across multiple devices. Mitkovski et al. [19] also utilized

a Pix2Pix GAN to emulate the P&S process for biometric applications.

To start with, and similarly to [19], we trained a Pix2Pix GAN [9] simulator. Training the Pix2Pix GAN, however, requires pixel-wise alignment of digital and P&S images for effective computation of the mean square error loss. To address this problem, we employed image alignment techniques during training. We also trained a CycleGAN P&S simulator, which supports unpaired image-to-image translation. In fact, CycleGAN does not necessitate paired images, thus greatly simplifying the preparation of the dataset.

A. Architecture of the simulators

Pix2Pix and CycleGAN have been extensively used to address various generative tasks. In our case, the objective of the Pix2Pix generator is to translate the input images from the digital to the P&S domain, while the discriminator is asked to distinguish between real P&S and digital pairs and their synthetic counterparts. The CycleGAN generators aim to translate images from the digital to the P&S domain and vice versa, ensuring cyclic consistency. With respect to classical CycleGAN training, we did not use the identity loss. In fact, printing a printed and scanned image again should not result in the identity operator, as the second P&S process would further degrade the image quality.

B. Dataset

To train the simulators, we used a dataset derived from the second version of the VIPPrint dataset [20]. This dataset consists of human face images printed with various modern color laser printers, each operating at different resolutions. Acquisition was performed using a TaskAlfa 3551 multifunctional scanner at 600×600 dpi resolution, and the images were saved using lossless compression. The size of the digital images is 1024×1024×3, while the P&S images are approximately 2036×2038×3, with slight variations of 5 to 10 pixels in both dimensions introduced during scanning. To align the resolutions of digital and P&S images, the digital images were upsampled to match the P&S image resolution. Our experiments focused on a subset of P&S images printed by one of the 12 printers in the VIPPrint dataset, specifically a Kyocera P5021 CDN Color Laser printer. We used a subset of 200 printed and scanned images from this printer for our experiments. To match the input size of the Pix2Pix and CycleGAN networks, we trained the networks on image patches extracted from 100 printed and scanned images along with their corresponding digital images. The patches were 256×256×3 in size and were extracted without pixel overlap. For Pix2Pix, we aligned the digital and printed and scanned patches using [21]. If alignment was challenging or significant pixel differences were detected, the corresponding patch was skipped. This approach yielded 4,678 aligned digital and P&S patches. In contrast, CycleGAN training utilized unaligned digital and P&S patches, leveraging the ability of CycleGAN's to handle unpaired image data. In total, 4,914 digital and P&S patches were used to train the CycleGAN simulator.

Fig. 1. Examples of digital and simulated P&S images with corresponding ground truth. The first column shows the digital image input, the second and third columns display P&S patches simulated by Pix2Pix GAN and CycleGAN respectively, and the last column shows the ground-truth P&S patches.

C. Training

The Pix2Pix GAN simulator was trained for 800 epochs using the Adam optimizer with parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 1×10^{-4}. The network utilized 64 filters and a Leaky ReLU activation function with a slope of 0.2, while the batch size was restricted to 1. For training CycleGAN, we used the same hyperparameters as the Pix2Pix GAN simulator over 600 epochs. Both the GAN adversarial loss and cyclic consistency loss weights were set to 10. After training, we assessed the performance of both simulators by inputting original digital patches. To introduce variability, we added Gaussian noise with zero mean and variance of 0.0625 to the digital images before feeding them to the simulator. This ensured that multiple inputs of the same digital image yielded slightly different simulated outputs, mimicking real-world variations when an image undergoes printing and scanning multiple times.

We evaluated the quality of the simulated images both visually (Fig. 1) and quantitatively using metrics such as the Structural Similarity Index (SSIM) and Fréchet Inception Distance (FID) (Table I). The SSIM scores are 0.84 for Pix2Pix GAN and 0.87 for CycleGAN, while the FID scores are 47 for Pix2Pix GAN and 45 for CycleGAN. As shown in Fig. 1, the images generated by the P&S simulators closely resemble the corresponding ground-truth images, demonstrating their effective learning of the distortions inherent in the P&S process.

TABLE I
SSIM AND FID BETWEEN SIMULATED AND REAL P&S IMAGES.

P&S Simulator	SSIM Score ↑	FID Score ↓
Pix2Pix GAN	0.84	47
CycleGAN	0.87	45

IV. GENERATION OF PHYSICAL ADVERSARIAL EXAMPLES

A. Threat Model

We consider an attack aiming to alter an image printed by a specific printer, *P*, in such a way that the printer source attribution model can no longer identify *P* as the source printer (untargeted attack) after the image is reprinted by *P*. The challenge is to ensure the attack's effectiveness even after the attacked image undergoes reprinting and scanning. The

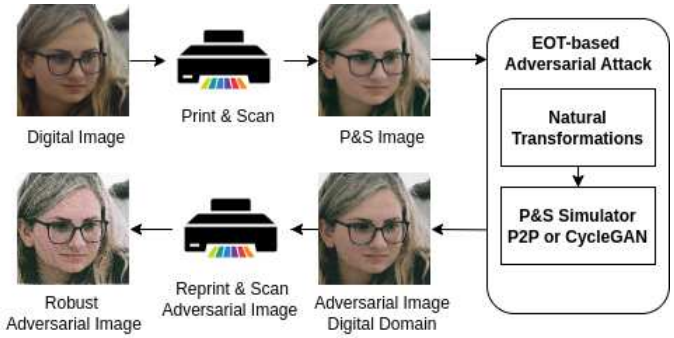

Fig. 2. Attack pipeline for the generation of robust adversarial examples.

attacker has white-box access to the source attribution model, including its weights and architecture. This allows the attacker to optimize and evaluate the adversarial examples in the digital domain before executing the physical-world attack by printing, scanning, and strategically placing the attacked images.

B. Targeted Model

The printer source attribution system targeted by our attack is the one described in [22]. This system, trained on the VIPPrint dataset [20], analyzes the 10 highest-energy 224×224×3 patches of the image and uses a majority voting decision for classification. Preliminary experiments in [22] showed that a basic reprinting black-box attack can deceive the original classifier. To enhance resilience against such attacks, the authors fine-tuned the model using a dataset of reprinted images, resulting in a more robust (hardened) source attribution model, which is the focus of our attack. Since the classifier operates on patches, the adversarial attacks are applied to 224×224×3 image patches. However, because the attack may slightly alter the energy of the patches, the classifier could potentially analyze different patches after the attack. Therefore, we decided to attack all patches in the image. This approach also prevents the introduction of visible discontinuities at patch borders. The success of the attack hinges on inducing sufficient patch misclassifications to misclassify the true printer as the most voted option. Our target is specifically the Kyocera-ecosys P5021cdn laser printer, identified as class #12 in the attribution system's multiclass classification.

C. Attack Pipeline

The attack pipeline (Fig. 2) begins with printing a digital image and applying an adversarial attack in the digital domain. To maintain the effectiveness of the adversarial perturbation after printing and scanning, we used EOT with P&S simulation. The adversarial digital image is then physically printed with the same printer. Finally, the source attribution model scans and analyzes the printed image to identify its origin.

1) Digital Domain Attack: Initially, we assessed the effectiveness of digital domain attacks (without EOT) in inducing misclassifications when the attacked image is subsequently printed and scanned. Adversarial examples were generated using a non-targeted version of I-FGSM [5] and C&W attack [12]. For I-FGSM, we set $\varepsilon = 0.03$, with a step size of 0.01 over 100 iterations. Similarly, for the C&W attack, we let $\varepsilon = 0.1$, with a binary search step size of 9 and a learning rate of 0.01 across 1000 iterations. These hyperparameters were selected to ensure effective attack coverage across most of the patches in the P&S image.

2) Physical Domain Attack: To create robust adversarial examples in the physical world, we integrated the I-FGSM and C&W attacks into an EOT framework, effectively addressing the domain shifts between digital and physical domains. EOT involves defining a pool of transformations T to simulate these shifts. The transformations used in our EOT attack are detailed in Table II, including their parameters, essential for replicating practical domain shifts. Additionally, we incorporated the Pix2Pix and CycleGAN P&S simulators within the transformation set. Results were averaged over 10 transformed samples to assess attack effectiveness. Through extensive experiments, we identified an optimal combination of transformations T (Table II) that consistently produce successful adversarial examples. Our setup includes rotation (2.0 to 10.0 degrees), zoom blur (factors between 1.05 and 1.10), and pixel shifts (5 pixels in all directions) with an inclusion probability of 100%. For color transformations, brightness deltas (10 to 40) and a fixed contrast factor of 0.3 are applied with 50% probability. Additionally, either CycleGAN or Pix2Pix GAN simulators are chosen with a probability of 50% to simulate P&S effects.

The attack algorithms within the EOT framework were configured with the following hyperparameters: for I-FGSM, $\varepsilon = 0.15$, a step size of 0.03, and 500 iterations were used; for C&W, we employed $\varepsilon = 0.15$, 9 binary search steps, a learning rate of 0.01, and 1000 iterations. When we incorporated the P&S simulators into EOT, I-FGSM utilized $\varepsilon = 0.4$, a step size of 0.07, and 500 iterations, while for C&W we used $\varepsilon = 0.4$, 9 binary search steps, a learning rate of 0.01, and 1000 iterations.

TABLE II
SET OF TRANSFORMATIONS USED IN THE EOT ATTACK.

Transformations	Parameter	Values	Probability
Brightness	brightness delta	[10, 40]	50%
Contrast	contrast factor	0.3	50%
Rotation	rotation angle	[2°, 10°]	100%
Zoom	zoom range	[1.05, 1.10]	100%
Shift of Pixels	# pixels (all directions)	5	100%
Pix2Pix P&S Simulator	-	-	50%
CycleGAN P&S Simulator	-	-	50%

V. EXPERIMENTAL RESULTS

To demonstrate the robustness of the hardened source attribution classifier against adversarial examples, we conducted experiments in both the digital and physical domains. Our study involved benchmarking various white-box attacks, including I-FGSM and C&W, both with and without EOT, and incorporating P&S simulators within the EOT transformations. These experiments were performed on a test set of 20 documents, with each document segmented into 81 patches,

(a) First P&S image (b) I-FGSM (c) I-FGSM(EOT) (d) I-FGSM(EOT+PS) (d) C&W (d) C&W(EOT) (d) C&W(EOT+PS)

Fig. 3. Examples of attacked images. From left to right, we show the original P&S image and the generated adversarial examples (after reprinting and scanning) using standard I-FGSM and C&W adversarial attacks, EOT with natural transformations, and EOT including P&S simulation.

totaling 1,620 attacked patches. To measure the strength of the perturbation introduced by the attacks, we computed the Peak Signal-to-Noise Ratio (PSNR) between the original and attacked patches, both before and after the reprinting and scanning process. PSNR calculations were focused only on successfully attacked patches. Additionally, we evaluated the ASR across all patches in the images and on the top 10 highest energy patches, which are generally more challenging to attack. After reprinting, the final classification is determined through majority voting on the results obtained from the top 10 highest energy patches of each document. To assess the overall robustness of the system, we also computed the ASR after the majority voting, where the printer with the largest number of votes among the top energy patches is selected.

The results of our experiments are reported in Table III. Analyzing the second column of the Table III, we observe that all attacks are highly effective when they are applied in the digital domain, achieving nearly 100% ASR. As expected, the attacks incorporating EOT, particularly those utilizing P&S simulation, exhibit lower PSNR values. The fourth column of the table, reports the effectiveness of the attacks in the physical domain, considering the ASR obtained on *all* patches after reprinting. For standard I-FGSM and C&W, the ASR decreases dramatically, while the application of EOT with natural transformations limits the ASR drop. Including the P&S simulators in the EOT transformations further improves the ASR to 80.49% for C&W and 87.83% for I-FGSM, which is a significant advantage with respect to EOT with natural transformations. The main advantage of including P&S simulation within EOT becomes apparent when we limit the analysis to the 10-highest energy blocks of each image. In this scenario, the ASR with natural EOT is only 33.5% for I-FGSM and 28% for C&W, while EOT with P&S simulation allows to attack 69% and 56.5% of the patches (for the attack to be successful it is necessary - though not sufficient - that at least 50% of the blocks are attacked). In the last column of the table, we report the ASR after majority voting on the 10-highest energy blocks, which is crucial for assessing the overall attack effectiveness on the entire image rather than on individual blocks. We observe that the ASR after majority voting drops to negligible values for standard I-FGSM and C&W attacks, showing only slight improvement with EOT using natural transformations [1]. However, when the P&S simulator is incorporated to EOT, the ASR significantly increases for both I-FGSM and C&W attacks. Specifically, the ASR for I-FGSM rises from 25% to 70%, and from 20% to 65% for C&W. These results highlight the effectiveness of incorporating the P&S simulator, given the complexity of creating adversarial examples that survive the reprinting process. Our experiments also suggest that patches with dark backgrounds tend to reintroduce stronger artifacts upon reprinting, thus requiring an excessive distortion.

TABLE III
EFFECTIVENESS OF VARIOUS ATTACKS IN BOTH THE DIGITAL AND PHYSICAL DOMAIN. ASR'S ARE AVERAGED ACROSS ALL PATCHES OF THE IMAGES, ON THE TOP 10 HIGHEST ENERGY PATCHES OF EACH IMAGE AND AFTER MAJORITY VOTING ON THE 10 HIGHEST ENERGY PATCHES.

Attack Method	ASR Digital	PSNR (dB)	ASR Printed **All Patches**	ASR Printed **Top10 Patches**	PSNR (dB)	ASR Printed **Majority Voting**
I-FGSM	100%	36.28	26.72%	15.5%	28.89	10%
I-FGSM (EOT)	96.41%	20.55	77.16%	33.5%	17.25	25%
I-FGSM (EOT+P&S)	100%	13.02	87.83%	69%	11.89	70%
CW	100%	34.25	21.48%	14%	25.53	10%
CW(EOT)	97.16%	19.86	63.70%	28%	16.96	20%
CW (EOT+P&S)	100%	12.28	80.49%	56.5%	11.18	65%

In Fig. 3, we present adversarial examples after reprinting, generated using various attacks. The images include the initial P&S image (the attack target), adversarial examples produced by standard attacks, EOT attacks with natural transformations, and EOT attacks incorporating P&S simulations. Comparing the initial P&S images to the reprinted adversarial examples generated by standard I-FGSM or C&W attacks we see that reprinting weakens the perturbation. The examples produced by I-FGSM(EOT+PS) and CW(EOT+PS) demonstrate the importance of the P&S simulation in creating robust adversarial examples that withstand reprinting.

[1]These results indirectly support the choice made in [22] to base the classification only on the highest energy patches.

VI. Concluding remarks

In our research, we addressed the challenges associated with generating robust adversarial examples against a source printer attribution system. We introduced a novel attack that integrates P&S simulations within the EOT framework. By employing Pix2Pix GAN and CycleGAN models, we developed two simulators that accurately replicate the P&S transformations. The integration of these simulators into the EOT framework significantly increased the ASR, demonstrating the method's effectiveness in producing adversarial examples that survive reprinting. Our work underscores the importance of physical domain adversarial attacks in AI security research and provides a foundation for future efforts to counteract such threats.

Future work will focus on developing defenses, such as adversarial training techniques that incorporate examples of images subjected to the proposed physical domain attack. We also plan to expand our simulators to address various image processing tasks under diverse environmental conditions. Additionally, we aim to develop advanced P&S simulators using diffusion models for enhanced realism and accuracy. Finally, we will explore vision transformers to challenge existing printer attribution systems.

Acknowledgment

This work was partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan, funded by the European Union - NextGenerationEU, and the FOSTERER project, funded by the Italian Ministry of University and Research (PRIN 2022 program, contract 202289RHHP).

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations*, 2014.

[2] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *CoRR*, vol. abs/2111.08429, 2021.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations*, 2015.

[5] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations*, 2017.

[6] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[7] Federal Trade Commission, "Consumer sentinel network databook," Tech. Rep., 2023.

[8] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning,*, 2018.

[9] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[10] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision, ICCV*, 2017.

[11] N. Purnekar, L. Abady, B. Tondi, and M. Barni, "Improving the robustness of synthetic images detection by means of print and scan augmentation," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec*, 2024.

[12] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017*, 2017.

[13] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face ID system," in *25th International Conference on Pattern Recognition*, 2020.

[14] J. Lu, H. Sibai, E. Fabry, and D. A. Forsyth, "NO need to worry about adversarial examples in object detection in autonomous vehicles," vol. abs/1707.03501, 2017.

[15] I. Evtimov, K. Eykholt, E. Fernandes, *et al.*, "Robust physical-world attacks on machine learning models," *CoRR*, vol. abs/1707.08945, 2017.

[16] S. T. K. Jan, J. Messou, Y. Lin, J. Huang, and G. Wang, "Connecting the digital and physical world: Improving the robustness of adversarial attacks," in *The Thirty-Third AAAI Conference on Artificial Intelligence,*, 2019.

[17] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability," *Comput. Vis. Image Underst.*, vol. 197-198, p. 102 988, 2020.

[18] M. Ferrara, A. Franco, and D. Maltoni, "Face morphing detection in the presence of printing/scanning and heterogeneous image sources," *IET Biometrics*, 2021.

[19] A. Mitkovski, J. Merkle, C. Rathgeb, *et al.*, "Simulation of print-scan transformations for face images based on conditional adversarial networks," in *BIOSIG*, 2020.

[20] A. Ferreira, E. Nowroozi, and M. Barni, "Vipprint: A large scale dataset of printed and scanned images for synthetic face images detection and source linking," *CoRR*, vol. abs/2102.06792, 2021.

[21] G. D. Evangelidis and E. Z. Psarakis, "Parametric image alignment using enhanced correlation coefficient maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1858–1865, 2008.

[22] A. Ferreira and M. Barni, "Attacking and defending printer source attribution classifiers in the physical domain," in *Pattern Recognition, Computer Vision, and Image Processing. ICPR*, 2022.