Attacking and Defending Printer Source Attribution Classifiers in the Physical Domain

Attacking and Defending Printer Source Attribution Classifiers in the Physical Domain

Anselmo Ferreira[1 [0000−0002−2196−7232]] and Mauro Barni[1 [0000−0002−7368−0866]]

Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy
anselmo.castelo@unisi.it,barni@dii.unisi.it

Abstract. The security of machine learning classifiers has received increasing attention in the last years. In forensic applications, guaranteeing the security of the tools investigators rely on is crucial, since the gathered evidence may be used to decide about the innocence or the guilt of a suspect. Several adversarial attacks were proposed to assess such security, with a few works focusing on transferring such attacks from the digital to the physical domain. In this work, we focus on physical domain attacks against source attribution of printed documents. We first show how a simple reprinting attack may be sufficient to fool a model trained on images that were printed and scanned only once. Then, we propose a hardened version of the classifier trained on the reprinted attacked images. Finally, we attack the hardened classifier with several attacks, including a new attack based on the Expectation Over Transformation approach, which finds the adversarial perturbations by simulating the physical transformations occurring when the image attacked in the digital domain is printed again. The results we got demonstrate a good capability of the hardened classifier to resist attacks carried out in the physical domain.

Keywords: Digital Image Forensics · Printer Source Attribution · Adversarial Attacks.

1 Introduction

Printed documents are everywhere. From printed advertisements and contracts to packages and anti-counterfeiting labels, a printer is always involved. However, the cheap access and high demand for new printing technologies raise many concerns about their misuse. For example, documents that could be considered a piece of evidence in a criminal investigation, such as illegal copies of documents, packaging of fake products, and terrorist plans, can be easily produced at anybody's home. Determining the provenance of printed documents, then, may be particularly important in several applications, such as anti-counterfeiting, forensics applications, and authentication of legal documents like statements, contracts, checks, among others. Indeed, according to a forecast from the International Chamber of Commerce, 3.7 trillion dollars and 5.4 million jobs will be

lost by 2022 [12] due to piracy. As another example related to piracy, according to the World Health Organization, almost 50% of the Malaria medications in Africa could be fake [37].

Current works on printed document forensics focus on two main tasks: the authentication of printed documents by pinpointing the ownership of a document (source attribution), and the description of copy-proof patterns that are distorted when an illegal copy of a printed authentication element is made (such as 2D barcodes). For the first task, which is the task of interest in this work, several papers have focused on identifying extrinsic artifacts such as noise, texture of printed patterns, banding, among others, and are usually divided in solutions focused on text documents [36, 34, 7, 18, 16, 17, 22], color documents [5, 38, 30, 20, 21, 10, 11] or both [8, 35, 3]. In all of these applications, methodologies based on artificial intelligence through Convolutional Neural Networks (CNNs) showed state-of-the-art performance [7, 18, 3, 10, 11].

Despite all these advancements, very little or no attention has been given to the evaluation of printed document forensics in adversarial conditions. A smart counterfeiter could, for example, use the ubiquitous existence of adversarial examples [23, 25, 24, 31, 2, 6, 32] to generate counterfeited labels that are judged as authentic by a CNN. In the same way, a criminal could modify a printed document to change the result of a source attribution procedure. For the specific domain of interest of this paper, some works have evaluated the transferability of adversarial attacks in the physical domain for computer vision applications. For example, Kurakin *et al.* [23] showed that printed and recaptured images could fool image classifiers by assuming that the images are presented in a given position. Other works have proven that variations of points of view do not impact the performance of attacks in the physical world [25, 24]. In [31], Sharif *et al.* attacked a face recognition system by proposing the printing of adversarial examples on a pair of eyeglass frames. Such an attack works by interactively looking for a perturbation that can fool the classifier, identified by optimizing a cross-entropy loss over a set of images that have already undergone geometric transformations typical of the recapture process. Athalye *et al.* [2] proposed an attack done by simulating synthetic transformations that can happen in the printing process of an image several times in the adversarial image construction. This is usually done to minimize the loss of an adversarial or target class (targeted attack) or maximize the loss of the real class (untargeted attack). Eykholt et al. [6] proposed an interesting physical domain attack to mislead stop sign classifiers. To reach their goal, they applied both synthetic and physical transformations and extended their work later to a general object recognition system [32]. Finally, the work by Zhang *et al.* [39] simulated the distortions a spoofed image is subject to when it is displayed on a smartphone screen to an anti-spoof authentication system. All the above-mentioned works have the same idea of modeling the possible physical world distortions that the image may be subject to during the attack optimization procedure. However, as will be discussed in the rest of this paper, applying such adversarial attacks against source attribution classifiers has some unique peculiarities that do not apply to other settings.

In this paper, we report our findings in fooling and defending printed documents source attribution classifier in the physical domain. To perform such a task, we first propose a simple black-box attack based on re-printing. Then we present a hardened version of the classifier, obtained by fine-tuning the original classifier using the attacked images obtained by reprinting. Finally, we evaluate the effectiveness of the hardened classifier against several white-box attacks, including a newly proposed method based on Expectation Over Transformation [2]. In summary, the contributions of this paper are:

1. We propose two adversarial attacks in the printed domain against source printer attribution classifiers: one of them is based on a simple, yet effective, black-box attack based on reprinting. The other is based on the Expectation Over Transformation strategy applied in a white-box setting.
2. We show how the simple black-box reprint attack can be enough to attack the printer source attribution classifier.
3. We introduce a new version of the source attribution classifier trained on examples attacked with the black-box attack, and demonstrate the effectiveness of such a defense against several white-box attacks, including the newly proposed method based on Expectation Over Transformation.

The rest of this paper is organized as follows: Section 2 discusses the threat model and the source attribution system targeted by our attacks. Section 3 presents our white and black-box attacks. Section 4 reports the results we got by attacking the original classifier with a simple rebroadcasting attack. In Section 5, we show the results of the experiments we run to validate the security of the hardened classifier. Finally, Section 6 concludes our work.

2 Threat Model

In this paper, we adopt the taxonomy introduced by Biggio and Roli [4], and already used in a previous work [39]. Before discussing the goal of the attack, we briefly review the system used for the printer source attribution problem targeted by our attacks.

The source attribution system, the attacks and defenses are based on the analysis and the datasets presented in [10, 9]. According to such works, focusing on an 8-class closed set scenario, the best source attribution is achieved by training a RESNET-50 [14] architecture. To apply the RESNET-50 CNN in the laser printer attribution task, the authors adopted classification over regions of interest (with further majority voting for classification), inspired by a previous work on rebroadcasting detection [1]. The source attribution is applied to high-energy regions detected after Canny filtering. Such energy is calculated on the Discrete Wavelet Transform domain and the top-10 highest energy $224\times224\times3$ patches of the documents are used for training and testing a CNN classifier. For the present work, we added to the dataset used in [10, 9] images taken from four new printers (already used in [11]). The images were also used to expand the VIPPrint dataset [10, 9] from 8 to 12 printers.

The threat model considered in this paper, which is illustrated in Figure 1, has two goals: (i) attack the high energy patches in such a way that these attacks remain in the print and scan domain; (ii) modify the patches in such a way to fool the source attribution classifier even after the majority voting scheme. As the classifier works on printed and scanned images, to perform the adversarial attack we print the digital images first, scan the images next and then we apply the adversarial attacks on them. After the attack, the attacked image is printed again, and the classifier is applied after re-scanning.

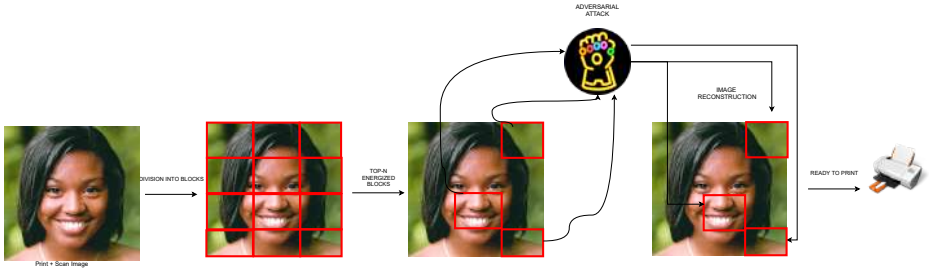


Fig. 1. The threat model considered in our work. We aim at modifying the high-energy patches using adversarial attacks to fool the source attribution system even after that the attacked image is re-printed and re-scanned.

In this work, we assume that the attacker has full access to the attacked system, representing a kind of worst-case assumption for the defender. This means that the attacker has not only access to the weights of the network, but he/she also knows which regions of interest will be used since the attacker also has access to the algorithm used to select the high-energy patches and will try to keep the same high energy patches after the attack. The adversarial replay attack consists of the construction of a scanned image with adversarial high-energy patches that could be scanned even with a scanner other than that used to train the classifier. The goal of the attack is to take an image printed with a certain printer $\mathcal{P}$ and modify it in such a way that the classifier does not recognize anymore that the image had been printed by $\mathcal{P}$. The challenge, here, is that the attack should be effective even after the attacked image has been printed again (by $\mathcal{P}$) and re-scanned. The attack we aim at is a purely exploratory attack [4], meaning that the attacker has no access to the training data used to train the classifier. Moreover, the attack is thought to work against an unattended system without human supervision. Finally, the printer $\mathcal{P}$ we focus our attack on is the Kyocera-ecosysp5021cdn laser printer, which is class #12 of our multiclass classification problem.

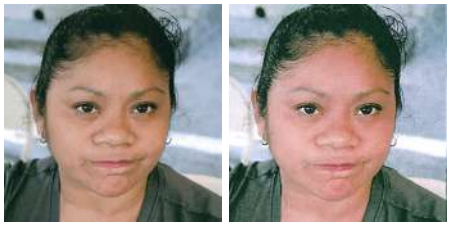
3 Proposed Attacks

In this section, we introduce the two attacks used throughout the paper. The first one is a black-box attack based on printing the document after it was already printed and scanned, and will be used to attack and defend the original model. The second attack is a white-box attack aimed at surviving the image distortions occurring between the first and second print. The last attack will be used to attack the hardened classifier fine-tuned with adversarial samples generated by the first method.

3.1 Double Rebroadcast Attack

The first adversarial attack we are considering relies on the experimental observation that a second print and scan process often fools the source attribution classifier with a significant margin. This observation contrasts with some related works, dealing with the effect of rebroadcasting in forensic applications. Zhang *et al.* [39], for example, applies Expectation Over Transformation to add a perturbation p that could fool a spoofing detection method once a spoofed image is displayed on a smartphone screen. The authors state that, without such a perturbation, the second rebroadcasted image could be easily detected as a spoof image by the anti-spoofing CNN. In our case, an image that has been printed and scanned twice can not be classified correctly, and thus a rebroadcasted document can be an adversarial attack against source printer attribution classifiers.

To support our claim, we show in Figures 2 and 3 some examples of images that are printed and scanned once and twice. We also show, in Figure 4, how the printing artifacts change in the second print by plotting the HH Discrete Wavelet Transform sub-bands of the first and second print. Such sub-bands were used by the work of Choi et al. [19] for printer attribution. Figure 4 shows that traces of the printer used for the second print are present when the image is printed twice (second and fourth columns of Figure 4), mixing artifacts and thus causing a classification error.



(a) Printed once (b) Re-printed

Fig. 2. First print (a) and second print (b) of the same image by considering a Kyocera Printer (printer #12) for first and second prints.

(a) Printed (b) Re-printed

Fig. 3. First print (a) with Samsung printer (printer #4) and (b) a reprinting of the first print using a Kyocera Printer (printer #12) for the second print.

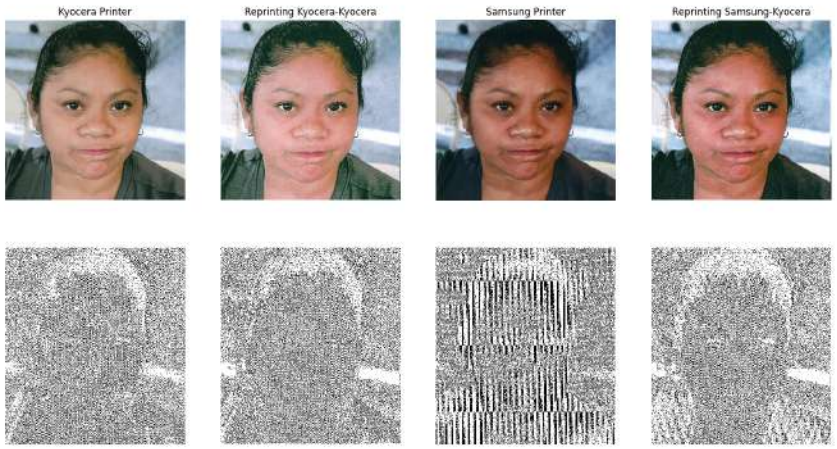


Fig. 4. Discrete Wavelet Transform HH subbands artifacts for images printed twice.

3.2 Expectation Over Transformation Attack

In this section, we present a white-box adversarial attack thought to be used against the hardened classifier fine-tuned on the reprinted images obtained by the black-box attack described in the previous subsection.

We start by formalizing the process whereby adversarial examples are generated. We may distinguish between targeted and untargeted attacks first. In the targeted case, given an input image I, we denote the target label of the adversarial example by $l_{adv} \neq l_{true}$, where l_{true} is the true label of the sample. We indicate with S_o the soft output of the neural network under attack. A targeted adversarial white-box attack aims at finding a minimal adversarial perturbation image I_δ solving the following optimization problem [39]:

$$\arg\min_{I_\delta \in \mathcal{I}} L(S_o(I + I_\delta), l_{adv}) + \lambda ||I_\delta||_p, \tag{1}$$

where $\mathcal{I}$ indicates the set of all possible perturbation images, L is the loss function of the neural network, $||\cdot||_p$ denotes the p-norm, and λ controls the strength of the distance penalty term $||I_\delta||_p$.

In the untargeted case, which is what we focus on in this paper, the optimization is rewritten as follows:

$$\arg\max_{I_\delta \in \mathcal{I}} L(S_o(I + I_\delta), l_{true}) + \lambda ||I_\delta||_p, \tag{2}$$

In the printed domain setting considered for our source attribution problem, the CNN is not fed directly with the digital image, but with its printed and scanned version. As the CNN works after the printing and scan operation, the attack is applied after the first print and scan process, then the adversarial image is re-printed and re-scanned. As will be discussed later, such an operation makes the source attribution attack scenario unique in terms of attack and defense strategies.

In order to better attack a printer source attribution system, the transformations the image suffers between the first and second print must be modeled during the attack. By denoting with T the set of distortions/transformations the attack must be robust to, the perturbation I_δ is found by optimizing the average loss over T as follows:

$$\arg\max_{I_\delta \in \mathcal{I}} E_{t \sim T}[L(S_o(t(I + I_\delta)), l_{true})] + \lambda ||I_\delta||_p, \tag{3}$$

In other words, for an adversarial perturbation I_δ to be successful in the untargeted case, Equation (3) states that it must maximize the mean loss of a true label after the application of several transformations to the adversarial image. In this way, after the attack, a random class will be given to the attacked image by the classifier (untargeted case). Such transformations are usually defined in such a way to simulate the distortions occurring when the adversarial image is rebroadcasted (in our case, re-printed and re-scanned). This formulation was first introduced by Athalye *et al.* [2] in the context of object recognition and later followed by Zhang *et al.* [39] to attack a face anti-spoofing classifier, and is commonly known as the Expectation Over Transformation (EOT) attack.

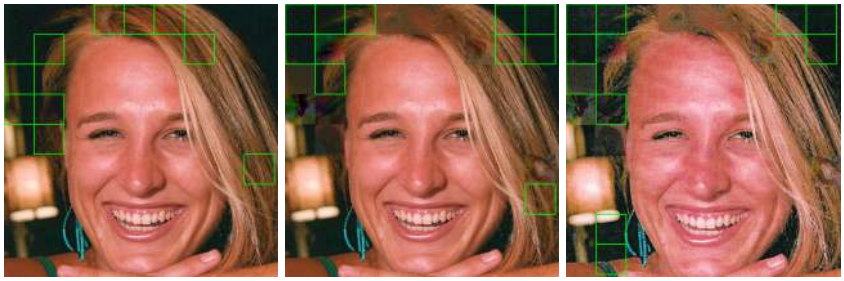
In the setting considered in [2], the degradation due to rebroadcasting is mainly due to geometric distortions, different color distributions, light reflections, and printing artifacts. Since in the printer attribution problem the scanning conditions are controlled with a fixed flatbed scanner, geometric distortion artifacts play a minor role. So, we focus on artifacts related to zoom, brightness change, and, most importantly, printing noise. The set of transformations used to implement our attack is listed in Table 1, together with the range of parameters we considered for each of them. For EOT, every t in Equation (3) is a composite transformation consisting of all the transformations in Table 1 applied randomly. The transformations in Table 1 were defined by visually inspecting

the transformations associated with reprinting, since the EOT-based method is expected to be applied in the physical domain, and hence the perturbation will have to survive a second print and scan process. The average loss is then computed over 100 versions of the to-be-attacked image, obtained by applying to it the composite transformations. To solve the minimization problem in Equation (3), we applied Stochastic Gradient Descent optimization (SGD).

Transformation	Range
Zoom	[-0.019,-0.029]
Brightness	[0.1,0.4]
Gaussian Noise (standard deviation)	[0.1,0.3]

Table 1. Transformations considered to simulate re-printing distortions for the Expectation Over Transformation based adversarial attack. For zoom effect, the interpolation used was bilinear and the values for brightness are the values that are randomly chosen and added to the image pixels in the normalized domain.

Finally, it is worth mentioning that, in our specific attack, we aim at keeping the position of the high energy patches unchanged after the attacked image is printed and scanned for the second time. To accomplish such a goal, a parameter called *strength*, or ϵ, is used to clip the values of the generated adversarial image. This is done by constructing an interval of accepted pixel values, so the original image is not degraded too much and thus the adversarial image will have an acceptable visual aspect. We show an example of the high energy patches in Figures 5 and 6 when considering a big and a small ϵ respectively. We found that from 70% to 80% of the high energy patches are unchanged in the attacked image when it is printed for the second time if a small ϵ is considered. In the experiments reported in the rest of the paper, we let $\epsilon = 0.01$.



(a) before attack (b) After attack (c) After printing

Fig. 5. Effect of a large strength value ϵ on the location of the high energy patches.

(a) Before Attack (b) After attack (c) After printing

Fig. 6. Effect of a small strength value ϵ on the location of high energy patches.

4 Effectiveness of the black-box attack against the original model

In this section, we report the results of the experiments we carried out to test the security of the original source attribution model against the simple black-box attack described in Section 3.1.

For the experiments described here and in the next section, we consider the dataset described by Ferreira *et al.* in [11]. The dataset consists of 200 printed faces, whose top-10 high-energy patches are used for printer attribution. Our attack focuses on one specific printer to which we had full access during our research: the Kyocera-ecosysp5021cdn, which we call printer #12. More specifically, we attack a model trained by considering 10 high-energy patches from 100 printed faces, printed by all 12 printers (10*100*12=12,000 images). For testing, we consider 100 printed faces from printer #12 and their top-10 high-energy patches to generate the adversarial examples.

To highlight the weakness of the original model against a simple reprint attack (hereafter referred to as `ReprintingAttack`), we considered three versions of the attack, varying the first and second printers as described below:

1. `ReprintingAttackv1`: we use the same printer (brand and model) for the first and second print: a Kyocera-ecosysp5021cdn.
2. `ReprintingAttackv2`: we use two different printers with the same brand, but with different models. We use a KyoceraTaskAlfa3551ci for the first print and a Kyocera-ecosysp5021cdn for the second one.
3. `ReprintingAttackv2`: we use two different printers with different brands and models. We use a Samsung-Multiexpress-X3280NR for the first print and a Kyocera-ecosysp5021cdn for the second one.

Figure 7 summarizes the errors of the classifier after majority voting on the top-10 high-energy patches.

The results reported in Figure 7 highlight the particularities of the printer source attribution problem in an adversarial setting. Simple black-box attacks like `ReprintingAttackv1`, `ReprintingAttackv2` and `ReprintingAttackv3` can

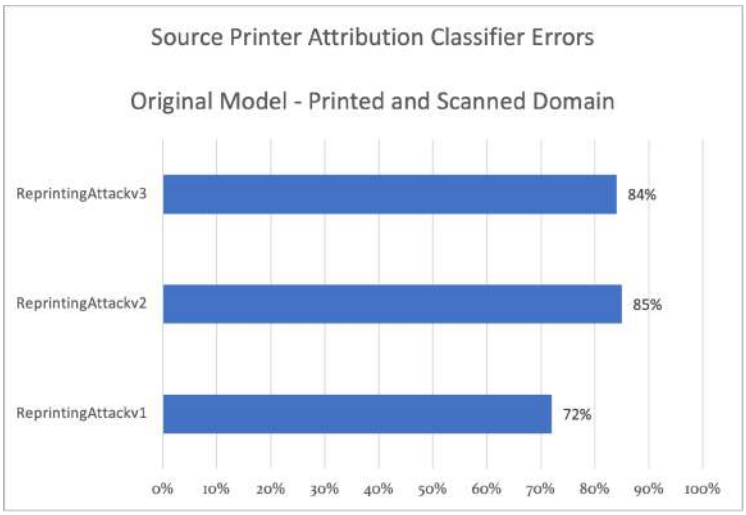


Fig. 7. Effectiveness of `ReprintingAttack` on the original classifier (error probabilities are reported as pecentages). The printer used for reprinting is always a Kyocera-ecosysp5021cdn printer. The results are given after majority voting on the high energy patches.

easily fool the original classifier, with error rates larger than 70%. Results at the patch level, not reported in the figure, confirm the effectiveness of the attack. In particular, `ReprintingAttackv2` and `ReprintingAttackv3`, using different printer sources for the first and second print, results in error rates around 50% and above 90% respectively. The effectiveness of these attacks motivates the development of a hardened classifier, trained on adversarial samples, as will be discussed in the next section.

5 Hardened source attribution model

We used the adversarial images obtained with the `ReprintingAttack` to harden the source attribution classifier. To train the new classifier, we fine-tune the original classifier by loading the original weights (previously found for the original printer source attribution). The fine-tuned classifier is trained on the original data plus the reprinting data (200 training images reprinted, being 100 of them with the same printer in the first and second prints, and the other 100 with different printers for the first and second prints, being the second printer always the Kyocera-ecosysp5021cdn). We establish as the source (or label) of reprinted data the printer that performed the second print. The architecture used to train the hardened model is the same RESNET-50 CNN as the original model, fine-tuned with SGD optimizer under 300 epochs and with a batch size of 32. An initial learning rate of 0.01 is defined and is reduced by a factor of $\sqrt{0.1}$ if there is a plateau in the validation accuracy curve for every 10 epochs. There is an early

stopping criterion that stops training if the validation accuracy does not change for 20 epochs. Finally, the best model considering such validation accuracy is saved.

In Table 2, we show the confusion matrix obtained by testing the hardened classifier in the absence of attacks. Not only the fine-tuned RESNET-50 model retains the good performance of the original model, but it also improves the original model's accuracy from 95,83% to 97,08%.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
#1	100,00%											
#2		82,00%	16,00%	2,00%								
#3		14,00%	85,00%	1,00%								
#4				100,00%								
#5					100,00%							
#6						98,00%	2,00%					
#7							100,00%					
#8								100,00%				
#9									100,00%			
#10										100,00%		
#11											100,00%	
#12												100,00%

Table 2. Confusion Matrix of a multiclass RESNET-50 source attribution model fine-tuned on reprinted images. The printer that we will focus on for our attack is highlighted in yellow in yellow last row and column.

5.1 Experimental analysis of the accuracy of the hardened classifier

To show the effectiveness of the hardened classifier against adversarial examples, we start evaluating its performance against adversarial attacks carried out in the digital domain. To do so, we benchmark a number of baseline white-box attacks against the hardened model in terms of performance metrics, selecting the best approaches to be applied in the physical domain. In particular, we focused on adversarial examples that remain effective even when they are converted from the normalized (floating point in the [0,1] range) domain back to the integer domain. Inspired by [33], we rely on the following metrics in both the normalized and integer domains:

1. **Error:** the inverse of accuracy, or the probability that a document is misclassified.
2. **L1 norm:** the mean of absolute pixel-wise differences between the original and the attacked images. Such a value is reported in percentage (0% means no variation, 100% means a variation from 0 to 255 in the pixel values).
3. **Linf norm:** the average maximum pixel-wise difference between the original and the attacked images (this is also given in percentage).
4. **PSNR:** the mean Peak Signal-to-Noise Ratio between the original and attacked images.
5. **%mod:** the average percentage of modified pixels when comparing the original and attacked images.

For the digital domain attacks, we considered the following attacks implemented in the Foolbox 2.0 library [28, 29]:

- the white box attack `L2FastGradientAttack`, also called the Fast Gradient Method (FGM) [13];
- the white box `LinfFastGradientAttack`, also called the Fast Gradient Sign Method (FGSM) [13];
- the white box `DeepfoolAttack` attack [27];
- `LinfPGDAttack`, which is the Projected Gradient Attack [26] using infinity norm;
- the `LinfBasicIterativeAttack`, which is the L-infinity norm Basic Iterative Method [23] and is built to work in the physical domain.

Tables 3 and 4 show results of the baseline white-box attacks in both normalized and integer domains before printing the adversarial attacks.

Approach	**Normalized Domain**				
	%Error	**Norm L1 (%)**	**Norm Linf (%)**	**PSNR**	**%mod**
`L2FastGradientAttack`	7.29	0.15	4.09	53.22	98.83
`LinfFastGradientAttack`	33.59	48.79	96.22	5.10	98.07
`LinfDeepFoolAttack`	33.9	10.21	10.60	22.09	96.98
`LinfPGDAttack`	100	5.41	11.03	24.54	99.26
`LinfBasicIterativeAttack`	100	1.82	3.64	34.67	99.27

Table 3. Results obtained by baseline white-box attacks against the hardened classifier in the digital normalized domain. Best results are highlighted in yellow and metrics are calculated patch-wise (before majority voting).

Approach	**Integer Domain**				
	%Error	**Norm L1 (%)**	**Norm Linf (%)**	**PSNR**	**%mod**
`L2FastGradientAttack`	6.89	0.12	4.08	53.53	26.73
`LinfFastGradientAttack`	33.4	48.79	96.22	5.10	98.07
`LinfDeepFoolAttack`	33.19	10.10	10.47	22.87	94.84
`LinfPGDAttack`	100	5.41	10.90	24.54	97.50
`LinfBasicIterativeAttack`	100	1.81	3.57	34.65	92.95

Table 4. Results obtained by baseline white-box attacks against the hardened classifier. Results are given after integer truncation and rounding of the normalized values. Best results are highlighted in yellow and metrics are calculated patch-wise (before majority voting).

According to Tables 3 and 4, `LinfPGD` and `LinfBasicIterative` are the most effective attacks, reducing the accuracy for printer #12 from 100% to 0%. Both of them generated adversarial samples by considering a small perturbation, highlighted by a small **Norm L1** and **Norm Linf**. For `LinfBasicIterativeAttack`,

the mean difference between the adversarial and original images in Table 4 is close to minimum (1.81%), and the maximum difference between them is also very small (3.57%), which means the adversarial perturbations are very small (and weak) in the digital domain. From Table 3 to 4, it can be seen that `LinfPGDAttack` and `LinfBasicIterativeAttack` do not suffer too much when passing from the normalized domain to the integer domain. The only obvious effect is a significant drop in the mean percentage of modified pixels (**%mod** metric), also affected by the truncation and rounding operations. As these approaches were successful in both the normalized and integer digital domains, we decide to also verify their effectiveness when applied in the physical domain (which means, printing the attacked printed images and converting them back in the digital format again).

Now we evaluate the effectiveness of such attacks against the hardened model, and in the presence of a further print and scan process (necessary to implement the attack in the physical domain). To do that, we selected as baseline adversarial attacks the best approaches found in the previous experiments (`LinfPGD` and `LinfBasicIterative`) and we also consider our proposed Expectation of Transformation based adversarial attack. In the sequel, we call our proposed attack as the `EOTReprintingAttack`.

Figure 8 shows the effectiveness of the adversarial attacks against the hardened detector, in the challenging setting imposed by physical domain attacks.

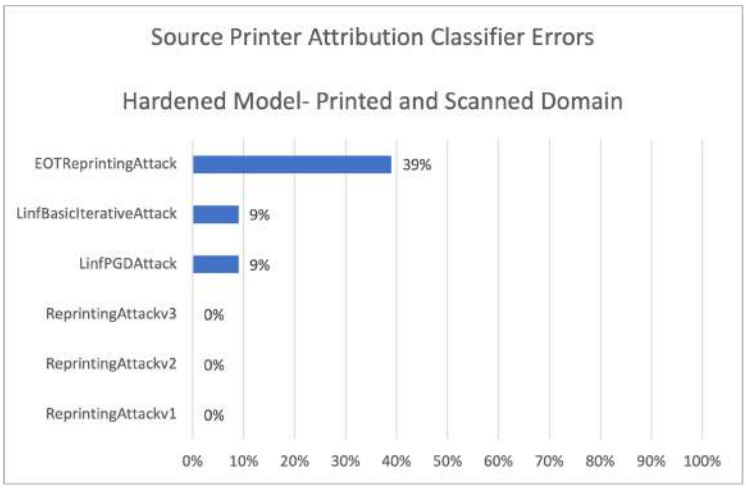


Fig. 8. Success rate of adversarial attacks against the hardened classifier when attacking the top-10 high energy patches. Results are averaged over 100 testing images. After the attack, the images are printed again with a Kyocera-ecosysp5021cdn printer. The results are given after majority voting.

To start with, the results in Figure 8 highlight the effectiveness of the hardened classifier against the simple rebroadcast attacks. For the baseline attacks, their effect drops dramatically due to reprinting and also the adversarial training of the to-be-attacked classifier, with an error rate below 10%. The proposed attack `EOTReprintingAttack`, being explicitly designed to cope with reprinting, exhibits a larger success rate equals to 39%. Still, the hardened classifier retains a good accuracy also in the presence of this powerful attack. We believe that the problem of `EOTReprintingAttack` not surviving the print and scan process is mainly due to the difficulty of simulating the artifact introduced by such a reprinting process,.

6 Conclusion

The security of any machine learning classifier, especially those interfacing directly to the physical domain, has gained importance as they have a substantial impact in several applications such as the reliability of self-driving cars, anti-spoofing systems, physical documents forensics, among many others. In spite of this interest, the security of image forensic tools operating on printed images source attribution has not been sufficiently studied. In this paper, we present some first steps towards this goal. We evaluated the security of a multiclass printer source classifier when it faces adversarial samples in both black box scenarios and white-box scenarios.

As an important finding of the research reported in this paper, we discovered that simple black-box attacks based on reprinting are often enough to attack an original source attribution classifier and thus can be used to harden it through adversarial training. We also proposed an attack based on Expectation Over Transformation to simulate reprinting artifacts in order to attack the hardened classifier. Despite these efforts, when facing the fine-tuned, hardened classifier, the performance of the attacks, including some baselines methods, is not satisfactory.

As future work, we aim to explore better ways to approximate the reprint and re-scan process when using Expectation Over Transformation attacks against protected models. One way that is under investigation is to rely on GANs like pix2pix networks [15] acting together with Expectation Over Transformation for the simulation of second print transformations, thus possibly improving the adversarial performance against adversarially-trained models.

Acknowledgements This research was funded by the European Union Marie Sklodowska-Curie project PrintOut (grant number 892757).

References

1. Agarwal, S., Fan, W., Farid, H.: A diverse large-scale dataset for evaluating rebroadcast attacks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1997–2001 (2018). https://doi.org/10.1109/ICASSP.2018.8462205

2. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Dy, J., Krause, A. (eds.) International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 284–293. PMLR (10–15 Jul 2018)
3. Bibi, M., Hamid, A., Moetesum, M., Siddiqi, I.: Document forgery detection using printer source identification—a text-independent approach. In: International Conference on Document Analysis and Recognition Workshops. vol. 8, pp. 7–12 (2019)
4. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition **84**, 317–331 (2018). https://doi.org/https://doi.org/10.1016/j.patcog.2018.07.023
5. Bulan, O., Mao, J., Sharma, G.: Geometric distortion signatures for printer identification. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1401–1404 (2009)
6. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1625–1634 (2018). https://doi.org/10.1109/CVPR.2018.00175
7. Ferreira, A., Bondi, L., Baroffio, L., Bestagini, P., Huang, J., dos Santos, J.A., Tubaro, S., Rocha, A.: Data-driven feature characterization techniques for laser printer attribution. IEEE Transactions on Information Forensics and Security **12**(8), 1860–1873 (Aug 2017). https://doi.org/10.1109/TIFS.2017.2692722
8. Ferreira, A., Navarro, L.C., Pinheiro, G., dos Santos, J.A., Rocha, A.: Laser printer attribution: Exploring new features and beyond. Forensic Science International **247**(0), 105 – 125 (2015)
9. Ferreira, A., Nowroozi, E., Barni, M.: VIPPrint: A Large Scale Dataset for Colored Printed Documents Authentication and Source Linking (Jan 2021). https://doi.org/10.5281/zenodo.4454971, Available at https://doi.org/10.5281/zenodo.4454971
10. Ferreira, A., Nowroozi, E., Barni, M.: Vipprint: Validating synthetic image detection and source linking methods on a large scale dataset of printed documents. MDPI Journal of Imaging **7**(3) (2021), https://www.mdpi.com/2313-433X/7/3/50
11. Ferreira, A., Purnekar, N., Barni, M.: Ensembling shallow siamese neural network architectures for printed documents verification in data-scarcity scenarios. IEEE Access **9**, 133924–133939 (2021). https://doi.org/10.1109/ACCESS.2021.3110297
12. Frontier Economics: The economic impacts of counterfeiting and piracy. Tech. rep., Frontier Economics (2016)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6572
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
15. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5967–5976. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.632
16. Joshi, S., Khanna, N.: Single classifier-based passive system for source printer classification using local texture features. IEEE Transactions on Information Forensics and Security **13**(7), 1603–1614 (2018)

17. Joshi, S., Khanna, N.: Source printer classification using printer specific local texture descriptor. IEEE Transactions on Information Forensics and Security **15**, 160–171 (2020)
18. Joshi, S., Lomba, M., Goyal, V., Khanna, N.: Augmented data and improved noise residual-based cnn for printer source identification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2002–2006 (2018)
19. Jung-Ho Choi, Dong-Hyuck Im, Hae-Yeoun Lee, Jun-Taek Oh, Jin-Ho Ryu, Heung-Kyu Lee: Color laser printer identification by analyzing statistical features on discrete wavelet transform. In: IEEE International Conference on Image Processing (ICIP). pp. 1505–1508 (2009)
20. Kim, D., Lee, H.: Color laser printer identification using photographed halftone images. In: European Signal Processing Conference (EUSIPCO). pp. 795–799 (2014)
21. Kim, D., Lee, H.: Colour laser printer identification using halftone texture fingerprint. Electronics Letters **51**(13), 981–983 (2015)
22. Kumar, M., Gupta, S., Mohan, N.: A computational approach for printed document forensics using surf and orb features. Soft Computing **24**(17), 13197–13208 (Sep 2020). https://doi.org/10.1007/s00500-020-04733-x, https://doi.org/10.1007/s00500-020-04733-x
23. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world (2016). https://doi.org/10.48550/ARXIV.1607.02533
24. Lu, J., Sibai, H., Fabry, E., Forsyth, D.: No need to worry about adversarial examples in object detection in autonomous vehicles (2017). https://doi.org/10.48550/ARXIV.1707.03501
25. Luo, Y., Boix, X., Roig, G., Poggio, T., Zhao, Q.: Foveation-based mechanisms alleviate adversarial examples (2015). https://doi.org/10.48550/ARXIV.1511.06292
26. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb
27. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2574–2582. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.282, https://doi.org/10.1109/CVPR.2016.282
28. Rauber, J., Brendel, W., Bethge, M.: Foolbox: A python toolbox to benchmark the robustness of machine learning models. In: Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning (2017), http://arxiv.org/abs/1707.04131
29. Rauber, J., Zimmermann, R., Bethge, M., Brendel, W.: Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. Journal of Open Source Software **5**(53), 2607 (2020). https://doi.org/10.21105/joss.02607, https://doi.org/10.21105/joss.02607
30. Ryu, S., Lee, H., Im, D., Choi, J., Lee, H.: Electrophotographic printer identification by halftone texture analysis. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1846–1849 (2010)
31. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. p. 1528–1540. Association for Computing Machinery, New York, NY, USA (2016)

32. Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramèr, F., Prakash, A., Kohno, T.: Physical adversarial examples for object detectors. In: USENIX Workshop on Offensive Technologies (WOOT 18). USENIX Association (2018)
33. Tondi, B.: Pixel-domain adversarial examples against cnn-based manipulation detectors. Electronics Letters **54** (08 2018). https://doi.org/10.1049/el.2018.6469
34. Tsai, M., Hsu, C., Yin, J., Yuadi, I.: Japanese character based printed source identification. In: IEEE International Symposium on Circuits and Systems (ISCAS). pp. 2800–2803 (2015)
35. Tsai, M., Yuadi, M., Tao, Y., Yin, J.: Source identification for printed documents. In: International Conference on Collaboration and Internet Computing (CIC). pp. 54–58 (2017)
36. Tsai, M.J., Yin, J.S., Yuadi, I., Liu, J.: Digital forensics of printed source identification for chinese characters. Multimedia Tools and Applications **73**(3), 2129–2155 (12 2014)
37. World Health Organization: A study on public health and socioeconomic impact of substandard and falsified medical products. Tech. rep., World Health Organization (2017)
38. Wu, H., Kong, X., Shang, S.: A printer forensics method using halftone dot arrangement model. In: IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP). pp. 861–865 (2015)
39. Zhang, B., Tondi, B., Barni, M.: Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability. Computer Vision and Image Understanding **197-198**, 102988 (2020)