Mobile authentication of copy detection patterns: how critical is to know fakes?

Olga Taran, Joakim Tutt, Taras Holotyak, Roman Chaban, Slavi Bonev and Slava Voloshynovskiy
Department of Computer Science, University of Geneva, Switzerland
{olga.taran, joakim.tutt, taras.holotyak, roman.chaban, slavi.bonev, svolos}@unige.ch

***Abstract*—Protection of physical objects against counterfeiting is an important task for the modern economies. In recent years, the high-quality counterfeits appear to be closer to originals thanks to the rapid advancement of digital technologies. To combat these counterfeits, an anti-counterfeiting technology based on hand-crafted randomness implemented in a form of copy detection patterns (CDP) is proposed enabling a link between the physical and digital worlds and being used in various brand protection applications. The modern mobile phone technologies make the verification process of CDP easier and available to the end customers. Besides a big interest and attractiveness, the CDP authentication based on the mobile phone imaging remains insufficiently studied. In this respect, in this paper we aim at investigating the CDP authentication under the real-life conditions with the codes printed on an industrial printer and enrolled via a modern mobile phone under the regular light conditions. The authentication aspects of the obtained CDP are investigated with respect to the four types of copy fakes. The impact of fakes' type used for training of authentication classifier is studied in two scenarios: *(i)* supervised binary classification under various assumptions about the fakes and *(ii)* one-class classification under unknown fakes. The obtained results show that the modern machine-learning approaches and the technical capacity of modern mobile phones allow to make the CDP authentication under unknown fakes feasible with respect to the considered types of fakes and code design.**

***Index Terms*—Copy detection patterns, printable graphical codes, copy fakes, supervised authentication, one-class classification**

I. Introduction

Nowadays, counterfeiting and piracy are among the main challanges for modern economy. Counterfeiting of medical supplies, food, cosmetics, mechanical parts and goods in general poses tremendous risks to public welfare and health, businesses and brand value reputation. At the same time, many traditional anti-counterfeiting technologies become quickly obsolete in view of rapid technological progress that offers a wide range of modern high-tech tools to the counterfeiters such as modern machine learning systems, high quality digital industrial printers and scanners. On the other hand, many new approaches to anti-counterfeiting appear thanks to the advancement of modern mobile technologies and machine learning algorithms.

In the recent years, the *printable graphical codes* (PGC) attracted a lot of attention as a link between the physical and digital worlds, which is of great interest for the internet of things, track and trace and brand protection applications. The anti-counterfeiting technology based on the PGC belongs to a family of hand-crafted physical unclonable functions (PUFs) [1]. Quite often, the PGC represent a union of the 2D bar codes that are clonable but have a semantic meaning and so-named *copy detection patterns* [2] (CDP) that are sensitive to the illegal copying. They might be combined in many different ways [3]–[5]. The PGC life cycle diagram is schematically shown in Fig. 1.

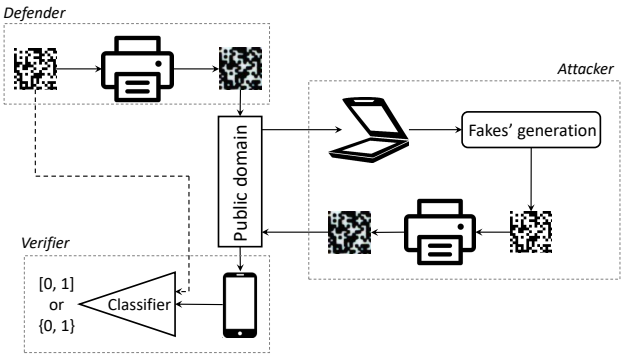


Fig. 1: General scheme of the CDP life cycle: *(i)* the generated digital templates are printed by a defender and go to the public domain; *(ii)* an attacker having an access to the publicly available codes can produce different types of fakes that are then also distributed in the public domain; *(iii)* a verifier digitizes the printed codes from the public domain and validates them via a parameterized classifier that might produce either a hard decision (fake/authentic $\sim$ 0/1) or a kind of a sort decision ranging from 0 to 1. The validation might be produced with or without taking the digital templates into account. For the defender-verifier pair the protection problem consists in the minimization of probability of error as a function of the CDP design, used printing and acquisition technologies and used classifier. For the attacker the goal is to maximize the probability of error as a function of the attack construction.

Up to our best knowledge, the CDP based authentication of the PGC under the conditions close to the real-life, where the codes are printed on an industrial printer and enrolled via the modern mobile phones has not been investigated in the prior art publications. In this respect, in this paper we perform this analysis for several types of copy attacks. Moreover, in

S. Voloshynovskiy is a corresponding author.

This research was partially funded by the Swiss National Science Foundation SNF No. 200021_182063.

TABLE I: An overview of the existing datasets of CDP: the datasets (1) and (2) are publicly available state-of-the-art datasets and the dataset (3) is created and investigated in the present paper.

#	Name	Digital templates	Printing	Acquisition	# of codes
(1)	CSGC [6]	size: 100×100 symbol size: 1×1	*Laser*, at 600 dpi: · Xerox Phaser 6500	*Scanner*: · Epson V850 Pro at 2400 ppi at 4800 ppi at 9600 ppi	digital: 950 original: 2850 fakes: 0 total: 3800
(2)	DP1C & DP1E [7]	size: 384×384 symbol size: 6×6	*Laser*, at 1200 dpi: · Samsung Xpress 430 · Lexmark CS310 *Inkjet*, at 1200 dpi: · Canon PIXMA iP7200 · HP OfficeJet Pro 8210	*Scanner*: · Canon 9000F at 1200 ppi · Epson V850 Pro at 1200 ppi	digital: 384 original: 3072 fakes: 3072 total: 6528
(3)	Indigo mobile	size: 330×330 symbol size: 5×5	*Industrial*, at 812 dpi: · HP Indigo 5500 DS	*Mobile phone*: · iPhone XS auto settings	digital: 300 original: 300 fakes: 1200 total: 1800

view of the fact that a particular type of fakes is unknown to the defender at the test stage, we propose and study an authentication system trained without knowledge of fakes and compare its performance with one that is trained with the complete knowledge of fakes.

Taking into account ethical and non-competition aspects of the considered problem with respect to several competitive technologies on the market, the investigation is performed on the CDP generated based on an open international standard ISO/IEC 16022 [8]. The main goal is to demonstrate a general approach applicable to the majority of CDP designed with the identical modulation principles rather than to investigate the authentication aspects of some particular CDP.

The main contributions of this paper are:

- A new dataset of CDP, which is produced on the industrial printing equipment HP Indigo 5500 DS and is acquired on the mobile phone iPhone XS under the regular light conditions.
- Investigation of the authentication aspects of CDP with respect to the typical copy fakes in a supervised and one-class classification setups.
- Analysis of the supervised and one-class classification of the CDP from the information theory point of view.

Notations: We use the following notations: $\mathbf{t} \in \{0,1\}^{m\times m}$ denotes an original digital template representing CDP; $\mathbf{x} \in \mathbb{R}^{m\times m}$ corresponds to the image of the original printed code, while $\mathbf{f} \in \mathbb{R}^{m\times m}$ is used to denote the image taken from a printed fake code; $\mathbf{y} \in \mathbb{R}^{m\times m}$ stands for a probe that might be either original or fake. $p_t(\mathbf{t})$ and $p_{\mathcal{D}}(\mathbf{x})$ correspond to the data distributions of the digital templates and original printed codes, respectively. A discriminator corresponding to the Kullback–Leibler divergence is denoted as $\mathcal{D}_{\mathrm{x}}$, where the subscript indicates the space to which this discriminator is applied to.

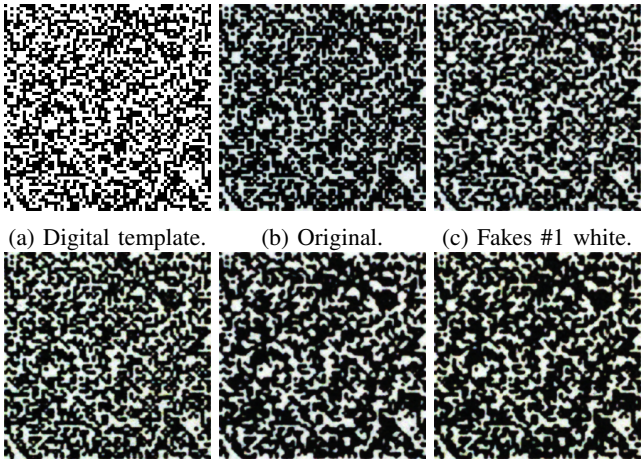
II. Indigo mobile dataset

Despite a big recent popularity of CDP, there are not so many public datasets available for investigation and reproducible research. The CDP dataset production is a time consuming and very costly process. Thus, the majority of the research experiments are performed either on synthetic data or on small private datasets. This also partially explains the lack of complete understanding about the clonability and performance of CDP under different classes of attacks.

Up to our best knowledge, there are only few public CDP datasets: *(i) DP0E* [9] and its extensions *DP1C* & *DP1E* [7] and *(ii) CSGC* [6]. The datesets' details are summarized in Table I. These state-of-the-art datasets were created to investigate the clonability aspects of CDP. Thus, the codes were printed on the desktop printers and enrolled by the scanners at the high resolution. At the same time, these conditions are not suitable to investigate the authentication of the CDP in the industrial settings, which is of great practical importance. In this respect, we present a new dataset, named *Indigo mobile* dataset that contains 300 distinct digital DataMatrix [8] templates $\mathbf{t} \in \{0,1\}^{330\times 330}$ with the symbols size of 5×5[1]. To simulate the real life scenario, the codes are printed on the industrial printer HP Indigo 5500 DS at the resolution 812 dpi.The acquisition of the printed codes is performed under regular room light using mobile phone iPhone XS under the automatic photo shooting settings at the resolution 12 Megapixels. The photos are taken in DNG format to avoid built-in mobile phone image post-processing. The final codes are converted to the RGB format based on the publicly available code [10][2]. Examples of digital template and corresponding enrolled printed original code are given in Fig. 2a and 2b, respectively. All codes are synchronized

[1]To ensure accurate symbol representation, each printed symbol should be represented by at least 3×3 pixels. The anti-counterfeiting is an important problem for the developing countries, where the relatively cheep phones with a low resolution (about 600 - 900 ppi) predominate. Taking into account the difference in the printing (812 dpi) and potential acquisition resolution (600 - 900 ppi) the symbols size should be about 4x4 or even 5x5.

[2]Despite the visually black and white nature of the CDP the authentication based on codes taken by the mobile phone in color mode is more efficient compared to the grayscale mode due to the different sensitivity of the color channels and corresponding degradation caused by converting a three-channels color image into a single-channel grayscale one.



(a) Digital template. (b) Original. (c) Fakes #1 white.

(d) Fakes #1 gray. (e) Fakes #2 white. (f) Fakes #2 gray.

Fig. 2: Examples of original and fake codes taken by a mobile phone from the Indigo mobile dataset.

using additional synhromarkers ensuring the correspondence between the digital templates and acquired codes.

As the most typical scenario for an unexperienced counterfeiter simulation we produce copy fakes based on the two standard copy machines (in a copy regime "text"): *(1)* RICOH MP C307 and *(2)* Samsung CLX-6220FX. The copy attacks based on advanced machine learning are addressed in [6], [7] and are not considered in this paper. The fakes are produced on the white and gray paper (80 g/m^2). Thus, for each original printed code we produce four fake codes:

1) *Fakes #1 white*: made by the copy machine *(1)* on the white paper.
2) *Fakes #1 gray*: made by the copy machine *(1)* on the gray paper.
3) *Fakes #2 white*: made by the copy machine *(2)* on the white paper.
4) *Fakes #2 gray*: made by the copy machine *(2)* on the gray paper.

The acquisition of the produced fakes is done in the same way on the same mobile phone under the same photo and light settings as for the original codes. Examples of the produced fakes are given in Fig. 2c - 2f. It is important to note that the fakes #1 visually closely resemble the original codes, while the fakes #2 have bigger dot gain and visually are more different from the original codes.

As it is summarized in Table I, the Indigo mobile dataset contains 1800 images of codes: *(i)* 300 distinct digital templates; *(ii)* 300 enrolled original printed codes and *(iii)* 1200 enrolled fake printed codes (300 originals × 4 type of fakes)[3].

For the simulation purposes, the Indigo mobile dataset was split into three subsets: *(1)* the training with 40% of data, *(2)* the validation with 10% of data and *(3)* 50% of data is used for the test. Taking into account a relatively small amount of data available for the training the next data augmentations are used: *(i)* the rotations on 90°, 180°and 270°; *(ii)* the gamma correction with variable function $(.)^\gamma$, where γ is the parameter of gamma correction.

[3]The Indigo mobile dataset is available at https://github.com/sip-group/snf-it-dis/tree/master/datasets/indigomobile.

III. Supervised classification

A. Theoretical analysis

In this paper we consider the supervised classification as a base-line scenario to validate the mobile phone based authentication of the CDP. The complete availability of fakes at the training stage in the supervised classification gives the defender an information advantage over the attacker.

To link the supervised classification problem with the considered one-class classification operating under the absence of fakes, we introduce a common theoretical basis based on an information-theoretic formulation. The problem of a supervised classifier training given the labeled data $\{\mathbf{x}_i, \mathbf{c}_i\}_{i=1}^N$ generated from a joint distribution $p(\mathbf{x}, \mathbf{c})$ is formulated as a training of a parameterized network $p_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})$ that is an approximation of $p(\mathbf{c}|\mathbf{x})$ originating from the chain rule decomposition $p(\mathbf{x}, \mathbf{c}) = p_{\mathcal{D}}(\mathbf{x})p(\mathbf{c}|\mathbf{x})$. The training of the network $p_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})$ is performed based on the maximisation of a mutual information $I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{C})$ between $\mathbf{x}$ and $\mathbf{c}$ via $p_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})$:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\, I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{C}), \tag{1}$$

that can be rewritten as:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathcal{L}_{\text{Supervised}}(\boldsymbol{\phi}), \tag{2}$$

where $\mathcal{L}_{\text{Supervised}}(\boldsymbol{\phi}) = -I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{C})$.

The mutual information in (1) is find as:

$$\begin{aligned} I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{C}) &\triangleq \mathbb{E}_{p(\mathbf{x},\mathbf{c})}\left[\log \frac{p_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})}{p_c(\mathbf{c})}\right] \\ &= \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{c})}\left[\log p_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x})\right]}_{\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}} - \underbrace{\mathbb{E}_{p_c(\mathbf{c})}\left[\log p_c(c)\right]}_{=\text{ constant}}, \end{aligned} \tag{3}$$

where $H(\mathbf{C}) = -\mathbb{E}_{p_c(\mathbf{c})}\left[\log p_c(c)\right]$ is an entropy of $\mathbf{c}$ and it is a constant that does not depend on $\boldsymbol{\phi}$.

The optimisation problem (2) reduces to:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, \mathcal{L}_{\text{Supervised}}(\boldsymbol{\phi}) = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} -\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}. \tag{4}$$

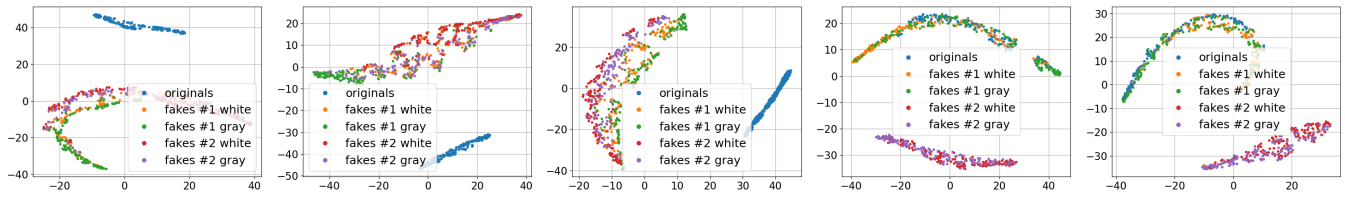
B. Empirical results

The main goal of this section is to investigate the influence of the different types of fakes on the authentication accuracy under the supervised classification. In this respect, the binary classifier was trained in five different setups[4] on the original codes and (1) all type of fakes, (2) fakes #1 white, (3) fakes #1 gray, (4) fakes #2 white, (5) fakes #2 gray.

[4]To avoid the bias in the choice of training and test data, in each setup the classifier was trained five times on the randomly permuted data. Each time the learning rate equaled to 1e-4, the batch size was 21 and the cross-entropy loss was used. The Adam was used as an optimizer. The gamma correction was performed with $\gamma \in [0.4, 1.3]$ with step 0.2. The more details about the used data augmentations are given in Section II.

TABLE II: The average (over five runs) classification error in % of the supervised binary classifier.

Setup	Original (P_{miss})	Fake #1 white (P_{fa})	Fake #1 gray (P_{fa})	Fake #2 white (P_{fa})	Fake # 2 gray (P_{fa})
(1) All fakes	0	0	0	0	0
(2) Fakes #1 white	0	0	0.14 ($\pm$0.32)	0	0
(3) Fakes #1 gray	0	0	0	0	0
(4) Fakes #2 white	0	99.43 ($\pm$0.32)	100	0	0
(5) Fakes # 2 gray	0	99.29 ($\pm$0.5)	99.86 ($\pm$0.32)	0	0



(a) (b) (c) (d) (e)

Fig. 3: The t-SNE visualization of the last layer before an activation function of the supervised binary classifier trained on the originals and (a) all type of fakes, (b) fakes #1 white, (c) fakes #1 gray, (d) fakes #2 white, (e) fakes #2 gray.

At the inference stage, the query sample $\mathbf{y}$, which might be either original $\mathbf{x}$ or one of the fakes $\mathbf{f}^k$, $k = 1, ..., 4$, is passed through the deterministic classifier $g_{\boldsymbol{\phi}}$ such that $p_{\boldsymbol{\phi}}(\mathbf{c}|\mathbf{x}) = \delta(\mathbf{c} - g_{\boldsymbol{\phi}}(\mathbf{x}))$ and $\delta(.)$ denotes the Dirac delta-function or simply $\mathbf{c} = g_{\boldsymbol{\phi}}(\mathbf{x})$. Herewith, $g_{\boldsymbol{\phi}}$ is trained with respect to the term $\mathcal{D}_{\mathrm{c\hat{c}}}$ in (4). The classification accuracy is evaluated with respect to the probability of miss P_{miss} and the probability of false acceptance P_{fa}:

$$\begin{cases} P_{fa} &= \Pr\{g_{\boldsymbol{\phi}}(\mathbf{Y}) = \mathbf{c}_1 \mid \mathcal{H}_0\}, \\ P_{miss} &= \Pr\{g_{\boldsymbol{\phi}}(\mathbf{Y}) \neq \mathbf{c}_1 \mid \mathcal{H}_1\}, \end{cases} \tag{5}$$

where $\mathbf{c}_1$ denotes a class of original codes, $\mathcal{H}_1$ corresponds to the hypothesis that the query $\mathbf{y}$ is an original code and $\mathcal{H}_0$ is the hypothesis that the query $\mathbf{y}$ is a fake code.

The obtained classification error is given in Table II. When all types of the fakes are available for training that corresponds to the setup (1) the classifier distinguishes the original codes with high accuracy even with respect to the most proximate fakes #1. This can also be seen in Fig. 3a, where the t-SNE [11] visualisations of the last layer before an activation function of corresponding trained classifier is shown. From this visualisation it is possible to observe two main clusters: the first one is formed by the original codes (blue cluster) and the second one is formed by the fakes (multi-color cluster). In the setups (2) and (3), where the classifier is trained only on the original codes and the fakes #1 (white - the setup (2) or gray - the setup (3)) the authentication error is low even with respect to the unseen fakes #2. The latent space visualisations presented in Fig. 3b and 3c are similar to the setup (1). In the setups (4) and (5), the situation is different: if during the training the classifier does not observe the high quality fakes (which correspond to the fakes #1 in our dataset) and is trained only on the fakes that are far away from the manifold of original data (in our dataset they correspond to fakes # 2), then the authentication of unseen high quality fakes becomes more complicated and less accurate as shown in Table II. The latent space visualisations (Fig. 3d and 3e) show that the fakes #1 form one cluster with the original codes and are almost certainly authenticated as the genuine codes.

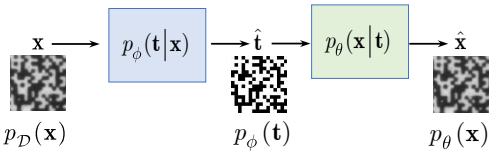


Fig. 4: General scheme of a deep model that aims at estimating the digital templates $\hat{\mathbf{t}}$ from the original printed codes $\mathbf{x}$ with the following mapping of the estimated digital templates $\hat{\mathbf{t}}$ back to the printed codes $\hat{\mathbf{x}}$.

IV. One-class classification

A. Theoretical analysis

The supervised classification scenario is an ideal case for the informed defender but it requires the knowledge of the corresponding fakes. With respect to the CDP authentication, the fakes' collection might be quite expensive and time consuming process. Moreover, taking into account the permanent improvement of technologies available for the fakes' production, it is very difficult to predict in advance what kind of fakes will be used by the attackers. In this respect, the one-class classification problem, where the authentication model is trained only on the original data disregarding the potential fakes, is of great practical importance.

In general case, it is possible to highlight the two main parts of the considered one-class classification system: *(1)* feature extraction and *(2)* one-class classifier. The one-class classifier by itself is an important part of the whole process. However, the main focus of this work is to find a set of reliable features

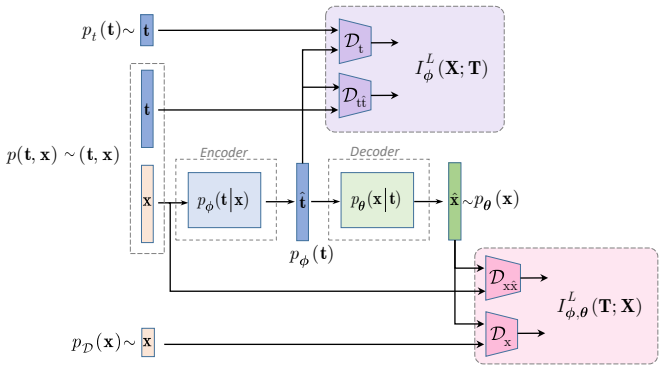


Fig. 5: The feature extraction based on the estimation of the digital templates $\hat{\mathbf{t}}$ via $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$ and $\mathcal{D}_{\mathrm{t}}$ and the printed codes $\hat{\mathbf{x}}$ via $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$ and $\mathcal{D}_{\mathrm{x}}$ terms.

that allow to distinguish between the original and fake codes even by using simple one-class classifiers. In this respect, we use the one-class support vector machine (OC-SVM) [12] as a one-class classifier model. Alternatively, one can use an one-class deep classifier [13].

As a feature extractor we investigate a deep auto-encoding model $\mathbf{x} \rightarrow \hat{\mathbf{t}} \rightarrow \hat{\mathbf{x}}$, where $\hat{\mathbf{t}}$ is considered as a latent space representation as shown in Fig. 4. The loss-function for the considered feature extracting system is:

$$\mathcal{L}_{\text{One-class}}(\boldsymbol{\phi}, \boldsymbol{\theta}) = -I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) - \beta I_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{T}; \mathbf{X}), \tag{6}$$

where β controls the relative importance of the two objectives.

The first mutual information term in (6) is defined as $I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})}{p_t(\mathbf{t})}\right]\right]$. According to [14], using the variational decomposition it can be lower bounded as $I_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) \geq I^L_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T})$, where:

$$\begin{aligned} I^L_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T}) \triangleq & \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})}\left[\log p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})\right]\right]}_{\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}} \\ & - \underbrace{D_{\mathrm{KL}}\left(p_t(\mathbf{t}) \| p_{\boldsymbol{\phi}}(\mathbf{t})\right)}_{\mathcal{D}_{\mathrm{t}}}, \end{aligned} \tag{7}$$

where $D_{\mathrm{KL}}\left(p_t(\mathbf{t}) \| p_{\boldsymbol{\phi}}(\mathbf{t})\right) = \mathbb{E}_{p_t(\mathbf{t})}\left[\log \frac{p_t(\mathbf{t})}{p_{\boldsymbol{\phi}}(\mathbf{t})}\right]$.

The second mutual information term in (6) determined as $I_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{T}; \mathbf{X}) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})}\left[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{t})}{p_{\mathcal{D}}(\mathbf{x})}\right]\right]$ can be decomposed and bounded in a way similar to the first term: $I_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{T}; \mathbf{X}) \geq I^L_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{T}; \mathbf{X})$, where:

$$\begin{aligned} I^L_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{T}; \mathbf{X}) \triangleq & \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{E}_{p_{\boldsymbol{\phi}}(\mathbf{t}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{t})\right]\right]}_{\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}} \\ & - \underbrace{D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right)}_{\mathcal{D}_{\mathrm{x}}}, \end{aligned} \tag{8}$$

where $D_{\mathrm{KL}}\left(p_{\mathcal{D}}(\mathbf{x}) \| p_{\boldsymbol{\theta}}(\mathbf{x})\right) = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[\log \frac{p_{\mathcal{D}}(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right]$.

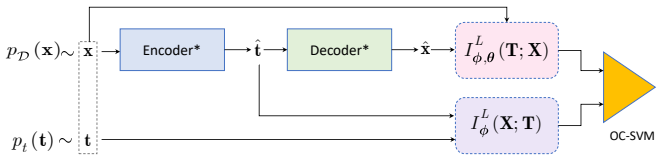


Fig. 6: The OC-SVM training procedure.

Combining the obtained decompositions the final optimization problem schematically shown in Fig. 5 is:

$$\begin{aligned} (\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) &= \underset{\boldsymbol{\phi},\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}^L_{\text{One-class}}(\boldsymbol{\phi}, \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\phi},\boldsymbol{\theta}}{\operatorname{argmin}} -(\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}} - \mathcal{D}_{\mathrm{t}}) - \beta(\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}} - \mathcal{D}_{\mathrm{x}}). \end{aligned} \tag{9}$$

For empirical evaluation of the theoretically obtained features' extractor we consider two basic scenarios of estimation of the digital templates $\hat{\mathbf{t}}$ and the printed codes $\hat{\mathbf{x}}$ based on:

- terms $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$ and $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$:

$$\mathcal{L}^1_{\text{One-class}}(\boldsymbol{\phi}, \boldsymbol{\theta}) = -\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}} - \beta\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}; \tag{10}$$

- terms $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$, $\mathcal{D}_{\mathrm{t}}$, $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$ and $\mathcal{D}_{\mathrm{x}}$:

$$\mathcal{L}^2_{\text{One-class}}(\boldsymbol{\phi}, \boldsymbol{\theta}) = -\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}} + \mathcal{D}_{\mathrm{t}} - \beta\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}} + \beta\mathcal{D}_{\mathrm{x}}. \tag{11}$$

B. Empirical results

The general schema of the OC-SVM training is illustrated in Fig. 6: the encoder and decoder parts of the feature extraction model shown in Fig. 5 are pre-trained and fixed (as indicated by a "*")[5] and the OC-SVM is trained on the different combinations of $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$ and $\mathcal{D}_{\mathrm{t}}$ terms' outputs that are the results of $I^L_{\boldsymbol{\phi}}(\mathbf{X}; \mathbf{T})$ decomposition and the $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$ and $\mathcal{D}_{\mathrm{x}}$ terms' outputs that are the results of $I^L_{\boldsymbol{\phi},\boldsymbol{\theta}}(\mathbf{T}; \mathbf{X})$ decomposition[6].

At the inference stage, the query sample $\mathbf{y}$, which might be either original code $\mathbf{x}$ or one of the fakes $\mathbf{f}^k$, $k = 1, ..., 4$, is at first passed through the feature extractor and then the corresponding feature vector is classified via pre-trained OC-SVM. The classification accuracy is evaluated with respect to the probability of miss P_{miss} and the probability of false acceptance P_{fa} given in (5).

From the obtained authentication results given in Table III one can note that the combination of the $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$ and $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$ terms' outputs is the most accurate feature vector among the considered ones for the OC-SVM. Moreover, it can be seen that the setup (4)-$\mathcal{L}^2_{\text{One-class}}$ produces the error that is two times

[5]The encoder and decoder models were trained with respect to the $\mathcal{D}_{\mathrm{t}\hat{\mathrm{t}}}$ and $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$ term correspondingly and were based on the U-Net architecture. The KL-divergence terms $\mathcal{D}_{\mathrm{t}}$ and $\mathcal{D}_{\mathrm{x}}$ were implemented in a form of density ratio estimator [15]. To avoid the bias in the choice of training and test data, the system was trained five times on the randomly shifted data. Each time the learning rate equaled to 1e-4, the batch of size 18 and the MSE loss were used. The Adam was used as an optimizer. The gamma correction was performed with $\gamma \in [0.5, 1.2]$ with step 0.1. The more details about the used data augmentations are given in Section II.

[6]The OC-SVM was trained to minimize the P_{miss} on the validation subset.

The python code for both supervised and one-class classification scenarios is available at https://github.com/taranO/Mobile-authentication-of-CDP.

TABLE III: The average (over five runs) authentication error in % of the one-class classification based on the OC-SVM.

Setup	Feature extractor	OC-SVM input	Original P_{miss}	Fake #1 white P_{fa}	Fake #1 gray P_{fa}	Fake #2 white P_{fa}	Fake #2 gray P_{fa}
(1)	$\mathcal{L}^1_{\text{One-class}}$	$\{\mathcal{D}_{\text{t}\hat{\text{t}}}, \mathcal{D}_{\text{x}\hat{\text{x}}}\}$	0.28 ($\pm$0.39)	0	0	0	0
(2)	$\mathcal{L}^2_{\text{One-class}}$	$\{\mathcal{D}_{\text{t}\hat{\text{t}}}, \mathcal{D}_{\text{t}}\}$	40.57 ($\pm$54.26)	0.57 ($\pm$0.59)	0.57 ($\pm$0.92)	0	0
(3)	$\mathcal{L}^2_{\text{One-class}}$	$\{\mathcal{D}_{\text{x}\hat{\text{x}}}, \mathcal{D}_{\text{x}}\}$	4.26 ($\pm$3.09)	0	0	2.55 ($\pm$3.08)	3.26 ($\pm$4.85)
(4)	$\mathcal{L}^2_{\text{One-class}}$	$\{\mathcal{D}_{\text{t}\hat{\text{t}}}, \mathcal{D}_{\text{x}\hat{\text{x}}}\}$	**0.14 ($\pm$0.32)**	**0**	**0**	**0**	**0**
(5)	$\mathcal{L}^2_{\text{One-class}}$	$\{\mathcal{D}_{\text{t}\hat{\text{t}}}, \mathcal{D}_{\text{t}}, \mathcal{D}_{\text{x}\hat{\text{x}}}, \mathcal{D}_{\text{x}}\}$	3.55 ($\pm$1.66)	0	0	0	0

smaller than the setup (1)-$\mathcal{L}_1$, although, the same combination of $\mathcal{D}_{\text{t}\hat{\text{t}}}$ and $\mathcal{D}_{\text{x}\hat{\text{x}}}$ terms' outputs are used for the OC-SVM input. It can be explained by the fact that the terms $\mathcal{D}_{\text{t}}$ and $\mathcal{D}_{\text{x}}$ in the $\mathcal{L}^2_{\text{One-class}}$ model play a role of learnable regularizers on the $\mathcal{D}_{\text{t}\hat{\text{t}}}$ and $\mathcal{D}_{\text{x}\hat{\text{x}}}$ terms correspondingly. That allows to make the estimations of the digital templates and printed codes more accurate for the original codes. This in turn leads to the decrease of the authentication error.

The obtained results demonstrate sufficiently accurate authentication of the CDP with respect to the typical copy attacks by a classifier trained only on the original codes without taking any fakes codes into account.

V. Conclusions

In the present work, we investigated the authentication aspects of the modern CDP produced under the conditions close to the real life: the codes are printed on the industrial printer and enrolled via the modern mobile phone under regular light conditions.

As a base-line we perform a theoretical and empirical evaluation of the supervised binary classification with defined training sets of fakes. The obtained results show that the fakes used for training play an important role: if classifier observes the high quality fakes at training that are close to manifold of the original codes it is capable to authenticate the unseen middle quality fakes with a high accuracy. In contrast, if during the training the classifier observes only the middle quality fakes it is not capable to authenticate unseen high quality fakes reliably.

To avoid a problem of search and acquisition of the suitable fakes for the training and theirs quick obsolescence in view of rapid technological progress, we investigated the one-class classification approach that does not require any fakes for the training. The investigated feature extraction system shows promising results even with respect to the one-class classification based on a simple OC-SVM model.

For the future work it is important to investigate the authentication aspects of the modern CDP with respect to the physical references and compare it with the considered authentication with respect to the digital templates. The combination of these two approaches is of great interest too. Finally, it is important to investigate the proposed methods to the CDP authentication with respect to the machine learning fakes.

References

[1] S. Voloshynovskiy, T. Holotyak, and P. Bas, "Physical object authentication: detection-theoretic comparison of natural and artificial randomness," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2029–2033.

[2] J. Picard, "Digital authentication with copy-detection patterns," in *Optical Security and Counterfeit Deterrence Techniques V*, vol. 5310. International Society for Optics and Photonics, 2004, pp. 176–183.

[3] J. Picard, P. Landry, and M. Bolay, "Counterfeit detection with qr codes," in *Proceedings of the 21st ACM Symposium on Document Engineering*, 2021, pp. 1–4.

[4] I. Tkachenko, W. Puech, O. Strauss, C. Destruel, and J.-M. Gaudin, "Printed document authentication using two level or code," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2149–2153.

[5] H. P. Nguyen, A. Delahaies, F. Retraint, D. H. Nguyen, M. Pic, and F. Morain-Nicolier, "A watermarking technique to secure printed qr codes using a statistical test," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2017, pp. 288–292.

[6] R. Yadav, I. Tkachenko, A. Trémeau, and T. Fournel, "Estimation of copy-sensitive codes using a neural approach," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2019, pp. 77–82.

[7] O. Taran, S. Bonev, T. Holotyak, and S. Voloshynovskiy, "Adversarial detection of counterfeited printable graphical codes: towards "adversarial games" in physical world," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] *ISO/IEC 16022: Information technology - Automatic identification and data capture techniques - Data Matrix bar code symbology specification*, 2006.

[9] O. Taran, S. Bonev, and S. Voloshynovskiy, "Clonability of anti-counterfeiting printable graphical codes: a machine learning approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, May 2019.

[10] "cv2.cvtcolor(): converts an image from one color space to another," https://docs.opencv.org/3.4/d8/d01/group__imgproc__color__conversions.html#ga397ae87e1288a81d2363b61574eb8cab.

[11] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *NIPS*, vol. 15. Citeseer, 2002, pp. 833–840.

[12] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class svm for learning in image retrieval," in *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, vol. 1. IEEE, 2001, pp. 34–37.

[13] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[14] S. Voloshynovskiy, O. Taran, M. Kondah, T. Holotyak, and D. Rezende, "Variational information bottleneck for semi-supervised classification," in *Entropy Journal special issue "Information Bottleneck: Theory and Applications in Deep Learning"*, August 2020.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.