Domain Generalization for Document Authentication against Practical Recapturing Attacks

Changsheng Chen, Senior Member, IEEE, Shuzheng Zhang, Student Member, IEEE, Fengbo Lan, Student Member, IEEE, Jiwu Huang, Fellow, IEEE

Abstract—Recapturing attack can be employed as a simple but effective anti-forensic tool for digital document images. Inspired by the document inspection process that compares a questioned document against a reference sample, we proposed a document recapture detection scheme by employing Siamese network to compare and extract distinct features in a recapture document image. The proposed algorithm takes advantages of both metric learning and image forensic techniques. Instead of adopting Euclidean distance-based loss function, we integrate the forensic similarity function with a triplet loss and a normalized softmax loss. After training with the proposed triplet selection strategy, the resulting feature embedding clusters the genuine samples near the reference while pushes the recaptured samples apart. In the experiment, we consider practical domain generalization problems, such as the variations in printing/imaging devices, substrates, recapturing channels, and document types. To evaluate the robustness of different approaches, we benchmark some popular off-the-shelf machine learning-based approaches, a state-of-the-art document image detection scheme, and the proposed schemes with different network backbones under various experimental protocols. Experimental results show that the proposed schemes with different network backbones have consistently outperformed the state-of-the-art approaches under different experimental settings. Specifically, under the most challenging scenario in our experiment, i.e., evaluation across different types of documents that produced by different devices, we have achieved less than 5.00% APCER (Attack Presentation Classification Error Rate) and 5.56% BPCER (Bona Fide Presentation Classification Error Rate) by the proposed network with ResNeXt101 backbone at 5% BPCER decision threshold.

Index Terms—Document Image, Recapture Detection, Deep Learning,

I. INTRODUCTION

There is urgent demand of online document authentication with increasing online services in the e-commerce and e-government applications. Authentication of digitally acquired documents image from hardcopy documents is an important forensic task with broad interest. However, the security loophole in the exiting authentication scheme has put our system at risk. As shown in Fig. 1 the text in an identity document (ID) image and a certificate image can be edited by some off-the-shelf image editing softwares (e.g., Adobe Photoshop) or some latest deep learning-based techniques [1].

To cover the forgery trace, the edited ID image can be re-acquired through a print-and-scan cycle. The resulting images in Fig. 1 are therefore more realistic than those edited in digital domain. It should be noted that the deep learning-based editing techniques proposed in [1] requires few human intervention, and the recapturing attack can be carried out by non-expert attackers using commodity printers and scanners. Thus, the document forgery attack together with recapturing operation has posed a new threat to our document authentication system.

Detection of document recapturing attack is the key step of identifying document from the forgery-and-recapture attack. However, recapturing detection of document image in practical applications is difficult since there are several variations in the recapturing process. First, the attacker may employ different imaging devices (such as a mobile/point-and-shot/professional camera), instead of a dedicated scanner. According to our preliminary study, a low-quality mobile imaging device introduce different capturing distortions compared to a high-quality scanner. Second, printing process with different printing techniques (such as inkjet, laserjet) with distinct halftoning patterns (such as screening, dithering) and substrates (such as office paper, high-quality glossy paper) should be considered. Third, the recapturing attack may be launched by completely different procedure. For instance, instead of the recapturing attack through the print-and-scan channel, the attacker may display the forged document on an LCD screen and re-acquire it with a camera, i.e., attack through the display-and-capture channel. Last but not least, there is a large variation of the document types in a practical document authentication system. It includes identity cards, certificates, contracts, etc. In summary, there is significant domain shift between samples collected under different settings. The main difficulty of document recapturing detection resides in generalizing the performance towards...
samples in different domains.

There are various research works on document image authentication. The existing techniques can be divided into the active and the passive categories. As an example of the active forensic techniques, the digital watermarking techniques can be applied to a document as protection against illegal alternations or re-acquisition. However, the active techniques demand controls over the document generation process, which limits its application to documents from various parties. In contrary, the passive techniques authenticate a document by the intrinsic features without the need to embed additional signals. For instance, a questioned document image can be examined through some inherent characteristics of the printing and acquisition process for tampering detection. However, the existing passive forgery detection techniques on digital images are not robust under the recapturing attack. Under such an attack, the tampered image of a given document is output with a printed and re-acquired with an imaging device to generate a recaptured version of the image. The recaptured document image has been through a complete image acquisition chain, and no post-processing (or forgery) after the acquisition steps. This image will be considered as an original copy by the existing passive tampering detection techniques. To fight against such attacks, the image recapturing detection has attracted global research attentions. However, most of the exiting recapture detection schemes focus on natural images and biometric images, only a few preliminary works have considered the recapturing attack on hardcopy documents. Worse still, as demonstrated in our experiment, the existing document recapturing detection schemes are not robust under domain variations in practical scenarios, such as printing/imaging devices, printing substrates, recapturing channels, and document types.

This paper focuses on the domain generalization problem for document authentication under recapturing attacks. Instead of tackling this problem in a general image authentication setting, we consider document recapturing detection in a verification setup with the aid of a genuine document image in the same template. Though requiring additional information in the authentication process, we want to emphasize two important points. First, such document authentication process that compares a questioned document with a reference is more intuitive according to the forensic process by a human expert. Second, the proposed scheme is practical for authenticating important documents (e.g., ID cards and patent certificates) with publicly available templates. With the above idea in mind, the proposed method takes advantage of the recent advancements in metric learning by constructing distinctive triplets (containing genuine, recaptured and reference document images) with considerations on the document contents and acquisition devices. The features of samples in a triplet are extracted by generic CNN backbones with shared weights. The distances between features in a triplet are then measured by an end-to-end trainable forensic similarity network, instead of employing the traditional Euclidean metric. The proposed loss integrates the forensic similarity function with a triplet loss and a normalized softmax loss. By optimizing the overall loss, a feature embedding space generated by the proposed scheme clusters the genuine samples while pushes the recaptured samples apart.

To evaluate the effectiveness of our approach, a database consists of 132 captured and 972 recaptured document images are collected. We consider a wide variety of experiment settings, including different printing/imaging devices, different printing substrates, different recapturing channels, and different types of documents. Experimental results show that the proposed schemes with different CNN backbones have consistently outperformed other approaches under various experimental settings. Specifically, under the most challenging scenario in our experiment, i.e., evaluation across different types of documents that produced by different devices, the proposed network has achieved less than 5.00% APCER (Attack Presentation Classification Error Rate) and 5.56% BPCER (Bona Fide Presentation Classification Error Rate) by the proposed network with ResNeXt101 backbone at 5% BPCER decision threshold.

The main contributions of this work can be summarized as follows.

- We propose a recaptured document authentication scheme by verification against a reference image using the Siamese network architecture. Instead of adopting the general Euclidean distance, the proposed network employs a trainable forensic similarity subnet to measure the distance in the embedding space.
- We take advantage of the recent advancements in metric learning and construct triplets with considerations on the document image contents, acquisition conditions, and resolutions. Distinctive features embedding space can therefore be learnt from triplets with various templates and different resolutions.
- We construct a practical document images database covering authentication scenarios with different printing/imaging devices, printing substrates, recapturing channels, and document types. The proposed scheme consistently demonstrates strong domain generalization performances under different experiment protocols.

The remaining of this paper is organized as follows. Section II reviews the related literatures on document recapturing detection and existing document databases. Section III introduces the proposed network architecture, forensic loss in the feature embedding space, triplet sampling strategy, and visualizes the feature embedding space. Section IV-A elaborates the data collection procedures of the captured and recaptured document images. Section IV-B compares the performance of different approaches under different settings, including intra-dataset, cross-dataset and in-the-wild experiments. Section V concludes this paper.

II. LITERATURE REVIEW
A. Related Works on Document Recapturing Detection

Existing image recapture detection schemes mainly focus on three types of media, i.e., biometric images, natural images, and document images. However, the biometric spoofing detection schemes aim at detecting the liveness of the inspected image. For example, in the problem of face spoofing detection,
researchers extract the forensic trace from rPPG signals (Remote photoplethysmography, i.e., heart pulse signal) and depth [8]. It should be noted that these features are not applicable to document images which are printed on flat paper surface.

The research of recapture detection algorithms focus on natural images starts as early as 2010 [9]. It shows that the original and recaptured images (from LCD screens) are not easily distinguishable by naked eye. Thongkamwitoon et al. detected the recaptured images based on edge blurriness and distortion which can be characterized by K-SVD dictionaries from a single image. Mahdian et al. presented a spoofing detection scheme with the periodic properties present in the LCD recaptured images. Li et al. proposed a hierarchical data-driven approach for image recapturing detection by extracting the intra-block information by CNN and capturing the inter-block dependency by recurrent neural network (RNN). Yang et al. proposed a Laplacian CNN model for small-size recaptured natural images from LCD screens, which is different from spoofing detection of hardcopy document images. Both the captured and recaptured versions of a hardcopy document are acquired from flat paper surfaces, which lacks the distinct differences between a 3D natural scene versus a flat and pixelated LCD screen. Worse still, to control production cost, a generic document is usually produced without employing special printing substrates and technologies.

There are only a few literatures on detecting spoofing document images. Shang et al. built a set of discriminate features from the characteristics of laser/inkjet printers and electrostatic copiers [6]. The features were then employed to distinguish documents produced by printing and copying devices. Over 90% classification accuracy has observed in a dataset collected with 10 laser printers, 6 inkjet printers and 9 copiers. Berenguel et al. adapted a recurrent comparator architecture with attention mechanism to detect forged banknote and ID document with recapturing operation [5]. It iteratively inspects different textual region of the document to distinguish a counterfeit document with an authentic one. Xu in Alibaba [7] proposed a recaptured detection scheme with the aid of built-in flash of the mobile phones. The document images are assumed to be acquired under a predetermined lighting condition such that the uncertainties in the capturing environment can be reduced. However, these works have not considered the domain variation in practical applications, such as different printing/imaging devices, printing substrates, recapturing channels, and document types.

B. Existing Document Databases

A publicly available database is important resource for benchmarking the performance of different approaches, but there is only a few document image datasets collected for the problem of recapturing detection. On the one hand, most recapture image datasets focus on natural images [9], [13] and biometric images [8]. On the other hand, most document image datasets were built for other applications, e.g., optical character recognition [14], [15], image classification [16], source device identification [17], and document forgery detection in digital domain (without recapturing) [18].

Recently, Shruiti et al. established a database of 14,500 rebroadcast images captured from hundreds of different devices through crowdsourcing [13]. The images are acquired by scanning a printed photo, rephotographing a displayed/printed photo. However, due to the uncontrolled nature of the sample collecting environment, the samples in this database are not reflecting the scenarios where an attacker tries to fool the document authentication system with careful operations. The quality of samples in this dataset vary significantly even in the same type of attack. Moreover, the content of the images mainly includes human subjects and natural scenes, which are very different from that processed by a document authentication system.

A dataset of 6615 banknote and 7550 legal ID document images from different countries is established in [15]. The dataset contains both original image and counterfeit document images produced by the scanning-printing operation. However, this dataset is not publicly available due to the sensitive content. Similar non-publicly available datasets have also been reported in [5], [19].

Given the literatures presented in Section II-A and II-B, it can be concluded that there are important limitations of the existing works. Firstly, the generality of existing document image recapture detection algorithms [5]–[7] towards various settings of the recapturing attack is remained to be investigated. According to our experimental results in Section IV-B, the state-of-the-art approach [5] is not performing well under some practical settings. Secondly, a publicly available dataset to evaluate different approaches under some representative settings is missing. Thus, in the following section, we propose a document authentication scheme with potential to identify recaptured document image under practical domain variations and establish a representative dataset (covering scenarios of different printing/imaging devices, printing substrates, recapturing channels, and document types) to benchmark the performance of different approaches.

III. THE PROPOSED DOCUMENT AUTHENTICATION SCHEME

In this work, we consider the document image authentication problem in a different perspective. Inspired by the fact that a human expert determines the authenticity of a questioned document usually by inspection side-by-side with a genuine document in the same template (or by checking some common features from a template), we believe that it is more intuitive to train a deep network to authenticate a document with the aid of a reference image, such as a high-quality authentic document image in the same template.

Thus, this work addresses the document authentication problem from the verification perspective. To allow comparing the questioned and reference document images through the network, we adopt a Siamese network architecture which takes
Fig. 2. Network architecture of the proposed recaptured document authentication scheme. The whole process can be divided into four stages, including triplet selection, feature extraction, similarity measurement, and forensic loss calculation.

two or more images as parallel inputs. The features of the input images are extracted by the network and compared in the latent domain. The authentication decision will be determined based on the distance between questioned and reference document images.

Given our analogy to the document inspection procedure of a human expert, vital parts of the proposed authentication scheme reside in extracting distinctive features, comparing features in the latent space and organizing the paired inputs. Thus, we will elaborate our network architecture for feature extraction, forensic loss in the feature embedding space, and sample pairing strategy in the following subsections.

**A. The Proposed Network Architecture**

In this part, we propose a deep learning architecture based on Siamese CNNs and metric learning. This is to allow multiple samples input to the network at the same time for mimicking the document inspection behavior of a human expert. The genuine, recaptured, and reference (high-quality genuine) document images in the same template are taken as input. As shown in Fig. 2, three patches (extracted from the positive, negative, and reference images, respectively), named a triplet, is input to the network simultaneously. The positive, negative, and reference patches are denoted as $x_{pi}$, $x_{ni}$, and $x_{ri}$, respectively, where $i$ is the index of triplet in a mini-batch. The recapturing traces, such as halftone and color degradation of the input samples are extracted by a set of generic CNNs with shared network weights. The extracted features are therefore denoted as $f(x_{pi})$, $f(x_{ni})$, and $f(x_{ri})$, respectively. The reference features are paired, respectively, with the features from positive and negative samples, yields two tuples $[f(x_{pi}), f(x_{ri})]$ and $[f(x_{ni}), f(x_{ri})]$. Each tuple is then evaluated by a trainable similarity network to compute distance between the two elements. By designing a customized loss function, the proposed deep model is trained to yield an embedding space that minimizes the distance between $[f(x_{pi}), f(x_{ri})]$ while maximizes the distance computed from $[f(x_{ni}), f(x_{ri})]$. Thus, a feature embedding space can be generated to cluster the positive samples near the reference, while pushes the negative samples apart. The proposed forensic loss in the feature embedding space is discussed in Section III-B.

Moreover, randomly choosing positive and negative samples from tens of thousands of possible combinations to form the triplets is neither efficient nor effective. The details of our triplet sampling strategy are elaborated in Section III-C.

In the testing procedure, we consider two scenarios depending on whether the template of the questioned document is included in the training data or not. First, if the template has been seen in the training data, the questioned document image is authenticated by comparing against both the positive and negative samples (of the same template) in the source domain. The first scenario shows the authentication process of some important documents, such as identity cards and birth certificate, which is in an official and publicly available template. Second, if the template is new to the training data, several pairs of positive and negative samples in the target domain would be required in the authentication process. This can be considered as a few shot learning scenario where only a few samples in the target domain are available. For both cases, the authenticity of a questioned document can be determined by the relative distances among the questioned, positive and negative samples. The distance metrics from three triplets (including three pairs of positive and negative samples, as well as a questioned sample) are evaluated to determine the authenticity of a questioned sample.

**B. Forensic Loss in the Feature Embedding Space**

General Euclidean distance has been widely adopted in comparing similarity between a pair of samples in the embedding space [20]. However, Euclidean distance merely cap-
tutes the equally weighted dimension-by-dimension distances between the two sets of features. The calculation of Euclidean distance assumes that different dimensions in a feature vector are uncorrelated and of the same variance [21]. Such distance metric may not be suitable to a document recapture detection task with large sample variation. For example, an unseen recaptured sample may have a smaller Euclidean distance with a genuine sample in the training set than that with a recaptured one. The design of a suitable distance metric in the embedding space is important.

Mayer and Stamm designed a CNN-based subnet in measuring the “forensic” similarity or distance between two samples in the embedding space [4]. This network is with a simple two-layer structure, including operations that map the input features to a high dimensional space, such as element-wise multiplication and concatenation. A similarity score is output at the final layer. A high similarity score indicates that the two input samples have a similar forensic trace, while a low similarity score implies that different forensic traces. However, the score is thresholded to 0 and 1 in computing the cross-entropy loss. Such binary labelling strategy is prone to over-fitting to the training samples (as demonstrated in our experiment).

To mitigate such issue, we propose to take advantages of both approaches. The forensic similarity function is incorporated into a set of general metric learning loss. Thus, the distances between two samples are measured by a trainable subnet, and the similarity score will be considered as a continuous distance metric.

In the following, we adopt a set of general metric learning loss in [22] with the forensic similarity function [4], yielding the proposed triplet similarity loss [23].

\[
\mathcal{L}_{ts} = \sum_{i=1}^{b} \left[ e^{ \alpha - S(f(x_{ri}), f(x_{pi})) + \gamma} - e^{ \alpha - S(f(x_{ri}), f(x_{ni})) + \gamma} \right] +
\]

\[
= e^{ \alpha} \sum_{i=1}^{b} \left[ e^{-S(f(x_{ri}), f(x_{pi}))} - e^{-S(f(x_{ri}), f(x_{ni}))} + \frac{\gamma}{\alpha} \right] +
\]

\[
= \sum_{i=1}^{b} \left[ e^{-S(f(x_{ri}), f(x_{pi}))} - e^{-S(f(x_{ri}), f(x_{ni}))} + \gamma \right] + (1)
\]

where \( b \) is the size of a mini-batch, \( \gamma \) is the a hyper-parameter that defines the safety margin between two sets of distance, \([\cdot]_{+}\) is the hinge function that ensures an operand only if it is non-negative, and \( S(\cdot, \cdot) \) evaluates the similarity between two samples with the deep network adopted from [4]. There are a few points to be noted in our triplet similarity loss defined in Eq. (1). First, the similarity score \( S(\cdot, \cdot) \) ranges from 0 to 1 indicates low and high similarity, respectively. Therefore, \( 1 - S(\cdot, \cdot) \) measures the disparity between a sample pair. Such disparity is then scaled by an exponential function which nonlinearly up-weights the samples with large distances. Second, Eq. (1) is achieved by eliminating a constant scaling factor \( \alpha \) from the loss function and by setting \( \gamma' = \gamma/\alpha \). After simplification, we see that the proposed triplet similarity loss encourages a large value of \( e^{-S(f(x_{ri}), f(x_{pi}))} \), i.e., a low similarity score between \( f(x_{ri}) \) and \( f(x_{pi}) \), while punishes a large value of \( e^{-S(f(x_{ri}), f(x_{ni}))} \). Thus, the positive samples will be clustered around the references, and the negative samples will be pushed away.

Besides, we also consider the relative distance between positive and negative samples with the normalized softmax loss [24].

\[
\mathcal{L}_{ns} = - \sum_{i=1}^{b} \log \left[ \frac{e^{S(f(x_{ri}), f(x_{pi}))}}{e^{S(f(x_{ri}), f(x_{pi}))} + e^{S(f(x_{ri}), f(x_{ni}))}} \right] + (a)
\]

\[
= - \sum_{i=1}^{b} \log \left[ \frac{e^{S(f(x_{ri}), f(x_{pi}))}}{e^{S(f(x_{ri}), f(x_{pi}))} + e^{S(f(x_{ri}), f(x_{ni}))}} \right] + (b)
\]

\[
= \sum_{i=1}^{b} \log \left[ 1 + \frac{e^{S(f(x_{ri}), f(x_{pi}))}}{e^{S(f(x_{ri}), f(x_{pi}))}} \right] + (2)
\]

where Eq. (2) is achieved by eliminating a constant scaling factor \( e^{-1} \) from both the nominator and denominator. The proposed normalized softmax loss is in a similar form with the proxy-NCA loss [24]. Eq. (2) rewrites the loss in a straightforward form. It indicates that, for each triplet, the minimum loss is \( \log(1+e^{-1}) \). The normalized softmax loss encourages a small \( S(f(x_{ri}), f(x_{pi})) \) and a large \( S(f(x_{ri}), f(x_{ni})) \) in a non-linear scale.

Overall, the proposed forensics loss can be written as

\[
\mathcal{L}_{fi} = \mathcal{L}_{ts} + \alpha \cdot \mathcal{L}_{ns} \quad (3)
\]

where \( \alpha \) is a hyper-parameter that specifies the weights of different loss components.

C. Triple Sampling Strategy

Employing all possible triplets to train our model is neither efficient nor effective [20]. As demonstrated in Section IV-B3 randomly sampling triplets from all samples introduces about 3.5% performance degradation compared to the proposed sampling strategy. Triple sampling strategy focus on mining representative triplets from all possible combination. Inspired by the workflow of a document forensic expert, who determines the authenticity of a signature by comparing the questioned specimen with a genuine one [25], we select a high-quality genuine document image as the reference sample in a triplet. To allow a more efficient training process, we generate triples from all samples by minimizing the variations, such as differences in document contents, acquisition conditions, and resolutions.

First, the triples are only generated among the document images from the same template (i.e., the same type of student ID card in our database) such that the image patches in each triple are with similar contents. It should be noted that the images in a triplet should be normalized to the same resolution to allow generation of patches with similar content, such as a face region as shown in Fig 2.

Second, only the samples with similar resolutions are chosen to form a triple. This is to avoid the quality loss in some heavily downsampled images. For example, if a scanned high resolution document image is compared with a captured low resolution one, heavy downsampling operation will be performed on the high resolution one to allow patch-to-patch comparison at the same location of two samples. However,
such downsampling operation introduces unnecessary difficulty in extracting useful forensic traces. Thus, the low- and high-resolution triplets are generated from samples collected by mobile phones and scanners, respectively, to avoid heavy downsampling.

Third, we also follow [4], [29] in selecting discriminative image patches to train our model. As shown in the triplet selection stage of Fig. 2, the discarded patches (highlighted in yellow) are those regions dominated by uniform backgrounds, and the patches with texts and graphics are more informative.

Finally, the semi-hard triplet mining strategy [27] is also adopted in our triple sampling process to allow a more efficient training process.

D. Visualization of Feature Embedding Space

As shown in Fig. 3 the resulting feature embedding separates the captured and recaptured samples in two different classes. To plot the t-SNE figures [28], we employ two different datasets captured and recaptured by different printing and imaging devices to train our deep model. These two training datasets are denoted as $D_1$ and $D_2$ according to the discussion in Section IV-A. The visualization shows how samples are separated in a two-dimensional space.

There are some interesting points to be noted. First, the captured and recaptured samples (blue and red solid circles) in $D_1$ and $D_2$ are clustered together though they are collected by two sets of different devices. It demonstrates that good generalization performance of the proposed deep model towards different devices, regardless of the training datasets. Second, most of the recaptured samples from an LCD screen (denoted as $D_3$ as discussed in Section IV-B3) are found to be in a single cluster marked by brown crosses. It indicates that the proposed scheme not only identifies the recaptured samples from an LCD screen, but also separates samples from different recapturing channels (i.e., print-and-capture and display-and-capture). Third, there are some purple crosses (high-quality recaptured samples with glossy photo paper, denoted as $D_4$ in Section IV-B3) among the blue and red solid circles (captured images in $D_1$ and $D_2$). Such phenomena does imply the difficulties in authenticating the high-quality recaptured samples.

IV. DATABASE AND EXPERIMENTAL RESULTS

A. Our Captured and Recaptured Document Image Database

To investigate the problem of document recapturing detection, a database consists of captured and recaptured document images is needed. We collect and share publicly a high-quality dataset for this problem. First and foremost, the content of our documents should be chosen carefully. Some legal documents (e.g., passport, ID card, certificate), which contain sensitive privacy, are not suitable to be shared in the public domain. Without loss of generality, student ID cards from 5 universities are chosen as the original document images in our experiment. As shown in Fig. 4 (a), these images are synthesized with Adobe CorelDRAW software according to the templates available in the public domain. The genuine documents are then manufactured by a specialized third-party on acrylic plastic to preserve the details in the graphical contents.

Our database contains two datasets. Dataset I ($D_1$) collects 84 genuine and 588 recaptured document images which are captured or recaptured, respectively, by $7 \times 7$ different combinations of devices. As shown in Fig. 5 the original document is printed by an authorized party to generate the genuine document, which are then scanned/recaptured to yield the captured document images. To collect the recaptured document images, the copied document images are print and re-acquired (by scanner or camera). Dataset II ($D_2$) follows the same data collection procedure but employs a different set of devices. It contains 48 genuine and 384 recaptured document images. As shown in Table I we have employed 4 phones, 3 scanners and 1 printer in collecting $D_1$, while 2 phones (including a high-quality camera phone, Oppo Reno with resolution of 48 mega-pixels (MP)), 2 scanners (including a high-end scanner, Epson V850 with optical resolution of...
TABLE I
The devices used for collecting dataset $D_1$ and $D_2$, which contains 672 samples and 432 samples, respectively.

| Set   | 1st Imaging Device | Printer                           | 2nd Imaging Device |
|-------|--------------------|-----------------------------------|--------------------|
| $D_1$ | Phone              | Scanner                           | Phone              |
|       | Xiaomi 8 (12 MP)   | HP OfficeJet 258 (4800 × 1200 DPI)| Xiaomi 8 (12 MP)   |
|       | RedMi Note 5 (12 MP) | Huawei P9 (12 MP)                  | RedMi Note 5 (12 MP) |
|       | Huawei P9 (12 MP)  | Apple iPhone 6 (8 MP)              | Huawei P9 (12 MP)  |
|       | Apple iPhone 6 (8 MP) | Brother DCP-1519 (1200 DPI)      | Apple iPhone 6 (8 MP) |
|       | Brother DCP-1519 (1200 DPI) | Epson V530 (1200 DPI)   | Brother DCP-1519 (1200 DPI) |
|       | Epson V530 (1200 DPI) | Beng K810 (1200 DPI)            | Epson V530 (1200 DPI) |
|       | Beng K810 (1200 DPI) |                                   | Beng K810 (1200 DPI) |
|       |                    |                                   |                    |
| $D_2$ | Phone              | Scanner                           | Phone              |
|       | Apple iPhone 6s (12 MP) | HP LaserJet m176n (2400 × 2400 DPI)| Apple iPhone 6s (12 MP) |
|       | Oppo Reno (48 MP)  | Epson L305 (2400 × 2400 DPI)     | Oppo Reno (48 MP)  |
|       |                    |                                   |                    |
|       | Epson V850 (3200 DPI) | Epson L305 (2400 × 2400 DPI)     |                    |
|       | HP LaserJet m176n (1200 DPI) | Epson V850 (3200 DPI) | HP LaserJet m176n (1200 DPI) |

Fig. 5. The block diagram of collecting genuine/captured document images and recaptured document images.

6400 dots per inch (DPI), and 2 printers (including a high-end printer, Epson L805 with resolution of 5760 × 1440 DPI) in gathering $D_2$. Thus, dataset II considers the recapturing attack with high-quality off-the-shelf devices.

To collect a high-quality image set, there are a few thumb rules in the data collection process of $D_1$ and $D_2$.

- **Camera Phones**: point perpendicular towards the document plane to avoid geometric distortion, set to the highest supported resolution, and the captured images are saved in JPEG format with the highest quality factor;
- **Environmental Light**: evenly illuminated by a lamp to avoid shadowing;
- **Scanners**: set to a resolution of 1200 DPI, except the Epson V850 remains at its default value of 3200 dpi, and the scanned output is saved in TIF format;
- **Printers**: set to color mode with the finest printing resolution;
- **Printing Substrate**: office paper with 120 g/m$^2$.

Two typical example of genuine and recaptured student ID images are shown in Fig. 4(b) and (c), respectively. It can be observed by comparing both images that the recapturing process introduces noise, blurring and color distortion.

B. Experimental Results

Based on the dataset presented in Sec. IV-A, we evaluate the performance of the proposed method and the benchmarking methods, including traditional handcrafted feature-based classifier (Local Binary Pattern + Support Vector Machine, LBP+SVM), generic CNN models (e.g., ResNet [30], ResNeXt [31], DenseNet [32]), and a state-of-the-art document recapturing detection scheme (the recurrent Comparator with attention models (CRC+Attn) [5]). Ablation studies on the forensic similarity (FS) subnet and triplet loss (TL) are also performed to demonstrate the advantages over [4] (a Siamese network with FS subnet for general image forensic) and [22] (a metric learning approach for face spoofing detection with similar triplet loss). To quantitatively compare different schemes, the performance metrics Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER), Equal Error Rate (EER), and Area Under the ROC Curve (AUC) are employed in our experiments.

The implementation details of different approaches are provided as follow.

**LBP+SVM**: We employ the LBP implementation in Matlab (version 2018b) with the default parameters to extract the LBP feature vector. The resulting feature vectors of image samples are then fed into the support vector machine (SVM) with linear and radial bias function (RBF) kernels.

**Generic CNN models**: We select 3 representative generic CNN models with different layer configurations, i.e., ResNet with 34, 50, 101 and 152 layers, ResNeXt with 50 and 101 layers, as well as DenseNet with 121, 169 and 201 layers. These CNN models are pre-trained with the ImageNet dataset and are fine-tuned with our document recapturing dataset to achieve a better performance. The original fully-connected (FC) layers of the pre-trained networks are replaced by two FC layers (with 256 nodes in the last layer) since our task only involves samples from 2 classes, instead of 1000 classes in the original ImageNet dataset. During our fine-tuning process, we freeze the weights of the shallow part of the network and use our database to retrain the weights of the FC layers. The input sample resolution of these networks is 224 × 224 pixels. The batch size is 128, and the model is trained with the Adam optimizer. The training period is 20 epochs with a learning rate of $1 \times 10^{-4}$ and the cross entropy loss. Our implementation is based on Pytorch 1.10 and Tensorflow 1.13.1.

**CRC+Attn**: We adopt the implementation provided in [5]. The recurrent comparator with attention model (CRC+Attn) is considered in our comparison since another configuration (i.e., recurrent comparator with co-attention, CRC+Co-Attn) suffers from the same limitation. Different from the data preprocessing in [5], pixel-precise alignment cannot be guaranteed in our dataset due to two reasons. First, the paired samples are not necessarily from the same ID image with the same content. More commonly, they are generated from ID...
TABLE II
EXPERIMENTAL RESULT IN DATASET D1 AND D2. THE SAMPLES IN EACH DATASET ARE DIVIDED IN AN 8:1:1 RATIO FOR THE TRAINING, VALIDATION AND TESTING SETS, RESPECTIVELY. THE BEST PERFORMANCE UNDER EACH CONDITION IS BOLD-FACED.

| Methods                  | D1                                      | D2                                      |
|--------------------------|-----------------------------------------|-----------------------------------------|
|                          | EER | AUC     | EER | AUC     | EER | AUC     |
| LBP+SVM (Linear)         | 23.08% | 0.8204 | 22.24% | 0.8380 |
| LBP+SVM (RBF)            | 25.00% | 0.8145 | 27.72% | 0.8056 |
| DenseNet121              | 0.01% | 1.0000 | 0.01% | 1.0000 |
| DenseNet169              | 0.01% | 1.0000 | 0.01% | 1.0000 |
| DenseNet201              | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNetX50                | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNetX101               | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet34                 | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet50                 | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet101                | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet152                | 0.01% | 1.0000 | 0.01% | 1.0000 |
| CRC+Attn [5]             | 24.47% | 0.8193 | 18.84% | 0.8938 |
| DenseNet121+FS           | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNetX101+FS            | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet50+FS              | 0.01% | 1.0000 | 0.01% | 1.0000 |
| DenseNet121+TL           | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNetX101+TL            | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet50+TL              | 0.01% | 1.0000 | 0.01% | 1.0000 |
| DenseNet121+Proposed     | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNetX101+Proposed      | 0.01% | 1.0000 | 0.01% | 1.0000 |
| ResNet50+Proposed        | 0.01% | 1.0000 | 0.01% | 1.0000 |

- CNN+FS: adopts Siamese network and the Forensic Similarity subnet [4];
- CNN+TL: employs Siamese network and the Triplet Loss in [22];
- CNN+Proposed: the proposed approach that combines both FS and TL.

images from different subjects. As shown in Fig. 2, some key user information and personal photo are different, and pixel-precise alignment is not possible. Second, some samples are acquired by mobile cameras which introduce lens, geometric, and surface distortions. It is not easy to align the details of two documents with pixel precision.

CNN+FS: We employ the forensic similarity (FS) subnet [4] to evaluate the distances between the features extracted by the CNN models. However, there are several distinctive differences between the architecture proposed in [4] and our implementation. First, the fine-tuned generic CNN backbones (ResNet, ResNetXt and DenseNet) are adopted in our implementation to extract the recapturing traces, such as halftone texture and color degradation, while the MISLNet CNN [35] is employed in [4] to extract image manipulation traces in digital domain. The network architecture in our implementation allows a fair and straightforward way of identifying the advantage of the FS subnet. Second, instead of using a fixed decision threshold on the similarity score (0.5, as shown in [4]), we determine the threshold by picking a value that optimizes the training accuracy. Third, the triplet sampling strategy in our implementation is different. The original implementation randomly chooses image patches with half of patch pairs chosen from the same and different camera models, respectively, while we generate triplets following the description in Sec. III-C by considering the uniqueness of our problem.

CNN+TL: We use the triplet loss (TL) proposed in [22] to train the Siamese-based CNNs. However, the original and our implementation of TL are different mainly in the two aspects. First, the triplets in [22] are generated through the semi-hard negative mining, while our implementation further considers the contents and resolutions of different document images as elaborated in Sec. III-C. Second, for fair comparison, we only utilize samples from the source domain to determine the authenticity of a questioned sample in the testing process, instead of computing a posteriori probability with samples in the target domain except for the results in Table VIII.

TABLE III
CROSS-DATASET EVALUATION IN D1 AND D2. D1 → D2 DENOTES TRAINING AND TESTING WITH DATASET D1 AND D2, RESPECTIVELY, WHILE D2 → D1 MEANS OTHERWISE. THE BEST PERFORMANCE FOR EACH BACKBONE NETWORK IS BOLD-FACED.

| Methods                  | D1 → D2                                      | D2 → D1                                      |
|--------------------------|----------------------------------------------|----------------------------------------------|
|                          | EER | AUC     | EER | AUC     | EER | AUC     |
| DenseNet121              | 14.40% | 0.9293 | 8.32% | 0.9731 |
| DenseNet169              | 17.39% | 0.9130 | 7.14% | 0.9822 |
| DenseNet201              | 19.64% | 0.8960 | 13.46% | 0.9363 |
| ResNetX50                | 21.73% | 0.8597 | 11.90% | 0.9487 |
| ResNetX101               | 10.85% | 0.9654 | 7.14% | 0.9795 |
| ResNet50                 | 17.28% | 0.8978 | 14.28% | 0.9333 |
| ResNet101                | 13.04% | 0.9440 | 16.66% | 0.9483 |
| ResNet152                | 24.34% | 0.8268 | 9.20% | 0.9756 |
| DenseNet121+FS           | 24.87% | 0.8757 | 4.75% | 0.9908 |
| ResNetX101+FS            | 2.61% | 0.9976 | 3.75% | 0.9955 |
| ResNet50+FS              | 3.93% | 0.9942 | 3.74% | 0.9940 |
| DenseNet121+TL           | 2.34% | 0.9976 | 3.40% | 0.9961 |
| DenseNet121+Proposed     | 14.51% | 0.9351 | 20.80% | 0.9024 |
| ResNetX101+Proposed      | 14.28% | 0.9361 | 17.14% | 0.9065 |
| ResNet50+Proposed        | 12.52% | 0.9395 | 12.55% | 0.9483 |
| DenseNet121+Proposed     | 3.40% | 0.9976 | 1.50% | 0.9985 |
| ResNetX101+Proposed      | 2.27% | 0.9988 | 1.47% | 0.9978 |
| ResNet50+Proposed        | 1.30% | 0.9979 | 4.00% | 0.9935 |
| DenseNet121+Proposed (R) | 6.51% | 0.9865 | 4.28% | 0.9930 |
| ResNetX101+Proposed (R)  | 6.86% | 0.9904 | 7.15% | 0.9814 |
| ResNet50+Proposed (R)    | 6.79% | 0.9891 | 5.67% | 0.9879 |

- CNN+Proposed (R): the proposed approach with random sampling strategy.

To investigate the challenges posed by recapturing attacks, we compare the performances of the proposed scheme with some benchmarking methods on our database under different experimental protocols.

- Intra-dataset Experiment: The training and testing images are from the same dataset (D1 or D2) acquired by the same set of devices;
- Cross-dataset Experiment: The training and testing images are from different datasets which involves different printing and imaging devices;
- In-the-wild Experiment: The training and testing images are from different printing substrates, different recapturing channels, and different types of documents.

1) Intra-dataset Experiment: In this experiment, the samples in each dataset are divided in an 8:1:1 ratio for training, validation, and testing sets, respectively. As shown in Table II...
performances of the generic recapture detection framework are satisfactory in the scenarios where training and testing data is sampled from the same subset. For example, most of the CNN-based schemes achieve EER = 0.01% and AUC = 1.0000 in the experiment conducted within $D_1$ and $D_2$, respectively. This experiment shows baseline performance of the CNN-based schemes under an ideal scenario.

However, there are several approaches which perform less satisfactorily. On the one hand, the LBP-based classifiers with default parameters achieve AUC less than 0.85. This is due to the limited descriptive power of the LBP features and the shallow structure of a SVM classifier. On the other hand, the AUCs of CRC+Attn are 0.81 and 0.89 in the Intra-dataset experiment of $D_1$ and $D_2$, respectively, which is a significant performance degradation compared to the original performance reported in [5]. Through investigation, we have found that this is because of the difference between the dataset in [5] and our datasets in Sec. IV-A. The triplet selection procedure in [5] selects the patches at the exact same position of the document, which requires precise alignment between different image samples. However, such precise alignment is not guaranteed in our datasets to reflect a more practical setting. As shown in Fig. 6 the mis-alignment among the input patches has introduced difficulties in identifying attention regions with the recurrent comparator which inspects the patches iteratively. It should be noted that such limitation is also observed in the CRC+Co-Attn approach [5].

2) Cross-dataset Experiment: We also evaluate the cross-dataset performances of different approaches (excluding LBP and CRC+Attn with poor performances in the intra-dataset experiment). The two datasets $D_1$ and $D_2$ are collected with two sets of different devices while the image content and acquisition environment (such as illumination conditions) remain the same. It can be observed in Table III that the recapture detection performances of the generic CNN-based schemes (ResNet, ResNeXt and DenseNet) degrade significantly when the training and testing data is inhomogeneous. The EERs of most generic CNN-based schemes are more than 10% (except DenseNet121, DenseNet169, ResNeXt101 and ResNet101 under $D_2 \rightarrow D_1$), which are over 10% increment compared to those of the cross-dataset experiment. Moreover, it is interesting to see that the performances under $D_2 \rightarrow D_1$ scenarios are better than those of $D_1 \rightarrow D_2$, except ResNet50. This is due to two reasons. First, the images dataset $D_2$ are collected with both laserjet and inkjet printers, while dataset $D_1$ only involves an inkjet printer. Second, the devices used in $D_2$ are of higher quality, such as Epson L805 with 5760×1440 DPI, Oppo Reno with 48 MP, and EPSO V850 with 3200 DPI. Comparing the two experiments ($D_1 \rightarrow D_2$ and $D_2 \rightarrow D_1$), it is easier to learn discriminative features from $D_2$ and to obtain better performances on a lower quality dataset $D_1$.

For the sake of comparison, only the CNN architecture with the best performance, i.e., ResNet50, ResNeXt101 and DenseNet121 are selected for our following comparison. Generally speaking, the CNN+FS and the proposed schemes worked much better than the other approaches. EERs of the CNN+FS and the proposed schemes are less than 5%. Meanwhile, the CNN+TL schemes do not have significant improvement over the generic CNN-based schemes. This can be explained by the Euclidean distance employed in [22] which is not a suitable metric for measuring the discrepancy between different samples in an cross-dataset setting. On contrary, the FS network [4] measures the sample distance with an end-to-end trainable network and leads to good performances for CNN+FS and the proposed schemes.

Moreover, to show the system performance under a practical setting, we evaluate the APCERs of different schemes while keeps the BPCERs at 1%, 5% and 10%, respectively. As shown in Table IV only the CNN+FS and the proposed schemes keep APCERs within 10% under pre-determined security levels set by different BPCERs. Similar to the results in Table III, the proposed approach outperforms CNN+FS under most conditions, except for DenseNet121 backbone under $D_1 \rightarrow D_2$ and ResNet50 backbone under $D_2 \rightarrow D_1$.

Last but not least, we performance an ablation study on the proposed triplet sampling strategy as described in Sec.
The random sampling strategy, denoted by ‘R’ in the last three rows of Table III is evaluated with the same number of triplets as the CNN+proposed configuration. On average, over 3.50% increment is observed in the EERs compared to the implementation with the proposed triplet sampling strategy.

3) In-the-wild Experiment: In this subsection, we investigate the performances of different schemes under some more practical scenarios where the training and testing images are from different printing substrates, different recapturing channels, and different types of documents.

- **Different Printing Substrates**
  Different printing substrates renders different printing distortions due to the different ink-substrate interaction characteristics [36]. As shown in Fig. 7, the ink spreading can be better controlled by employing a high-quality printing substrate. Compared to the recaptured sample in Fig. 4-(c), these high-quality samples in Fig. 7 contain less noise. To evaluate the robustness of the proposed scheme under different printing substrates, we employ iPhone 6s, HP LaserJet M176n and Epson L805 (belong to the set of devices in D2) and glossy photo paper to collect 80 high-quality recaptured samples. This dataset is denoted as D3.

In this part, the pre-trained models obtained from Table IV are tested against the recaptured samples in D3. It should be noted that all the samples in D3 are recaptured samples, and the performance metric for this case is APCER. According to different training sets (D1 or D2) used, this experiment includes scenarios with same or different recapturing devices, i.e., D2 → D3 or D1 → D3. In general, the scenarios with different devices are more challenging than those with the same devices.

It can be seen in Table V that the proposed schemes outperform CNN+FS and CNN+TL under all settings. Let us focus on the performance under BPCER=5% with the same devices, the APCERs of proposed schemes are less than 5% which shows at least 75% improvement over those of CNN+TL and CNN+TS.

- **Different Recapturing Channels**
  As is well-known that the recapturing attack can also be carried out with the display-and-capture channel [9]. Such recapturing attack with an LCD screen is very different from the reprinting process considered in our datasets D1 and D2. Comparing Fig. 4-(c) and Fig. 8, we see that there is less distortion in noise (mainly introduced by printing process) but heavier distortion in color in the recaptured samples from an LCD screen. In this part, we consider a cross domain testing. The training samples are collected by printers and scanners, while those in the testing set are recaptured from an LCD screen. To facilitate our study, we collect a high-quality dataset following the setting in [37] to avoid visible moiré
TABLE VI
Evaluation on different recapturing channels of the deep models trained by $D_1$ and $D_2$, respectively. The decision thresholds are set by BPCER = 1%, 5%, and 10% during the training of models presented in Table IV. APCER is employed as the performance metric. The best performance for each backbone network is bold-faced.

| Methods                  | Same Imaging Devices ($D_1 \rightarrow D_2$) | Different Imaging Devices ($D_2 \rightarrow D_4$) |
|--------------------------|---------------------------------------------|----------------------------------------------|
|                          | BPCER=1% BPCER=5% BPCER=10%                | BPCER=1% BPCER=5% BPCER=10%                |
| DenseNet121              | 100.00% 100.00% 95.38%                     | 100.00% 94.05% 82.74%                      |
| ResNeXt101               | 100.00% 100.00% 100.00%                    | 98.17% 98.17% 86.13%                       |
| ResNet50                 | 100.00% 100.00% 89.40%                     | 100.00% 100.00% 100.00%                    |
| DenseNet121+FS           | 24.33% 12.50% 9.52%                        | 36.90% 18.45% 13.69%                       |
| ResNeXt101+FS            | 25.21% 16.07% 9.52%                        | 25.00% 15.48% 0.59%                        |
| ResNet50+FS              | 28.57% 17.26% 10.12%                       | 30.36% 16.67% 12.50%                       |
| DenseNet121+TL           | 13.48% 11.90% 8.93%                        | 23.60% 16.67% 11.31%                       |
| ResNeXt101+TL            | 13.45% 10.00% 12.50%                       | 19.64% 13.69% 0.83%                        |
| ResNet50+TL              | 13.11% 10.71% 5.36%                        | 19.05% 12.50% 10.12%                       |
| DenseNet121+Proposed     | 12.50% 7.74% 5.36%                         | 0.60% 0.60% 0.00%                          |
| ResNeXt101+Proposed      | 3.57% 2.38% 0.60%                          | 1.19% 1.19% 0.00%                          |
| ResNet50+Proposed        | 8.93% 5.95% 5.36%                          | 2.98% 2.38% 2.38%                          |

Fig. 9. Samples of certificate document images. (a) a genuine certificate image captured by Oppo Reno. (b) a recaptured certificate image collected by Oppo Reno (both first and second imaging process) and Epson L805.

pattern. In this part, we use two smart phones (Huawei P9 and iPhone 6) in dataset $D_1$ since this dataset contains more genuine document images (to be displayed in the recapturing process) than $D_2$. Totally, 168 recaptured images from an LCD screen from ThinkPad X1 with 1920×1080 pixels resolution is gathered. This dataset is denoted as $D_4$. According to different training set ($D_1$ or $D_2$) used, this experiment includes scenarios with same or different recapturing devices, i.e., $D_1 \rightarrow D_2$ or $D_2 \rightarrow D_4$.

It can be seen in Table VI that the generic CNN models (DenseNet121, ResNeXt101 and ResNet50) are not able to generalize towards different recapturing channels. This is because the recapturing artifacts in printing and displaying devices are different, but the generic CNN models have only been trained to classify the artifacts generated in the printing process, such as halftone patterns. For the samples recaptured from displaying devices, no halftone pattern can be observed due to the limited display resolution. Therefore, over 94% of the recaptured samples are mistaken as genuine samples under 0.05 BPCER by the generic CNN models.

The proposed schemes work well under this cross domain testing with APCERs lower than 0.10 except the case with DenseNet121 under 0.01 BPCER and the same imaging devices $D_1 \rightarrow D_4$. It is not straightforward that the proposed schemes perform much better under the scenarios with different imaging devices than those with the same imaging devices. By analyzing the results in both classes, we found that the error samples are mainly high-quality recapturing images which are similar to the samples reprinted by a high-quality printer (Epson L805) in $D_2$. Thus, pre-training the model with $D_2$ leads to a higher performance among the high-quality recapturing samples.

- **Different Types of Documents**

A practical document authentication system processes different types of document images, such as ID card, certificate and other legal documents. It is necessary to evaluate a cross domain scenarios where the training and testing samples from different types of documents. Worth noting that, this scenario not only considers different document contents, but also include the variations in different printing substrates (e.g., the acrylic plastic of ID cards and office paper of other documents), as well as different printing techniques. To facilitate this study, we have collected some document images with the same printing (Epson L805 and HP LaserJet m176n) and imaging devices (HP LaserJet m176n, Epson V850, and Oppo Reno) as $D_2$. Samples of certificate document images are shown in Fig. 9. This dataset, denoted as $D_5$, contains 36 and 60 genuine and recaptured certificate images in 6 different templates. According to different training set ($D_1$ or $D_2$) used, this experiment includes scenarios with same or different devices, i.e., $D_2 \rightarrow D_3$ or $D_1 \rightarrow D_5$.

This experiment is conducted exclusively on the proposed scheme since the generic CNNs, CNN+FS and CNN+TL have failed in the less challenging tests on printing substrates and recapturing channels. During the testing process, the template of a questioned document is new to the training data, three pairs of positive and negative samples in the target domain are employed. This can be considered as a few shot learning setting where only a few samples in the target domain are available.

As shown in Table VII, evaluations of the pre-trained models show that the BPCERs of different backbone networks are higher than 0.5 in most case. It reflects a high authentication error rate for the genuine document images. This is due to the triplet loss and forensic similarity network are not generalizable across different types of documents. More specifically, the genuine documents of ID cards and certificates are manufactured by different printing techniques and different printing substrates. The two types of genuine documents are therefore with distinctive features.
To further evaluate this scenario, we select 6 triplets (18 images) in two different certificate templates to fine-tune our models. To balance the samples (i.e., the number of patches) from different types of devices, 2 and 4 triplets from scanners and mobile cameras are selected. After fine-tuning with 6 triplets (18.75% of the samples in D₃), we observe significant performance improvement in APCERs. Specifically, the APCER of the proposed scheme with ResNet50 backbone (at 5% APCER) has decreased from 50% and 70% to 11.11% and 11.11%, respectively, for the experimental scenarios with the same and different devices.

V. CONCLUSION

In this work, we have proposed a recaptured document detection scheme by taking advantages of both metric learning and image forensics techniques. Experimental results have demonstrated that the proposed scheme has good generalization performance under variations of printing (inkjet and laser printers)/imaging (high-quality scanners and low-quality camera phones) devices, substrates (general office paper and glossy photo paper) and recapturing channels (print-and-scan and display-and-capture). The proposed scheme is simple but competitive. Moreover, we have presented the first publicly available image dataset for document recapturing problem covering practical but challenging situations.

In the future, the remaining issues in the scenario of different document types (i.e., the third bullet in Sec. IV-B3) should be further investigated in two aspects. First, the samples in target domain used to fine-tune this model should be reduced or be eliminated to improve the practicality of our approach. Second, it is also beneficial to study a scenario where the reference image in the target domain is not available.

REFERENCES

[1] L. Zhao, C. Chen, and J. Huang, “Deep Learning-based Forging Attack on Document Images,” arXiv preprint arXiv:2102.00653, 2021.

[2] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, Digital Watermarking and Steganography, Morgan Kaufmann, 2007.

[3] P.-J. Chiang, N. Khanna, A. K. Mikkilineni, M. V. O. Sgovia, S. Suh, J. P. Allebach, G. T.-C. Chiu, and E. J. Delp, “Printer and Scanner Forensics,” IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 72–83, 2009.

[4] O. Mayer and M. C. Stamm, “Forensic Similarity for Digital Images,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1331–1346, 2020.

[5] A. Berenguel, O. R. Terrades, J. L. Canet, and C. C. Morales, “Recurrent Comparator with Attention Models to Detect Counterfeit Documents,” in International Conference on Document Analysis and Recognition, 2019, pp. 1332–1337.

[6] S. Shang, N. Memon, and X. Kong, “Detecting Documents Forged by Printing and Copying,” EURASIP Journal on Advances in Signal Processing, vol. 2014, no. 1, p. 140, 2014.

[7] W. Xu and J. Chen, “Method, Apparatus and Electronic Equipment for Recaptured Image Detection,” Jan. 2020, CN Patent 110705532A.

[8] W. Sun, Y. Song, C. Chen, J. Huang, and A. C. Kot, “Face Spoofing Detection based on Local Ternary Label Supervision in Fully Convolutional Networks,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3181–3196, 2020.

[9] H. Cao and A. C. Kot, “Identification of Recaptured Photographs on LCD Screens,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 1790–1793.

[10] P. Yang, R. Ni, and Y. Zhao, “Recapture Image Forensics based on Laplacian Convolutional Neural Networks,” in Proceedings of International Workshop on Digital Watermarking, 2016, pp. 119–128.

[11] P. Yang, R. Li, R. Ni, and Y. Zhao, “Recaptured Image Forensics based on Quality Aware and Histogram Feature,” in Proceedings of International Workshop on Digital Watermarking, 2017, pp. 31-41.

[12] A. Anjum and S. Islam, “Recapture Detection Technique based on Edge-types by Analysing High-frequency Components in Digital Images Acquired through LCD Screens,” Multimedia Tools and Applications, pp. 1–21, 2019.

[13] S. Agarwal, W. Fan, and H. Farid, “A Diverse Large-scale Dataset for Evaluating Rebroadcast Attacks,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 1997–2001.

[14] V. Arlazarov, K. Bulatov, T. Chernov, and V. Arlazarov, “MIDV-500: a Dataset for Identity Document Analysis and Recognition on Mobile Devices in Video Stream,” Computer Optics, vol. 43, no. 5, pp. 818–824, 2019.

[15] K. Bulatov, D. Matalov, and V. V. Arlazarov, “MIDV-2019: Challenges of the Modern Mobile-based Document OCR,” in Proceedings of International Conference on Machine Vision, vol. 11433, International Society for Optics and Photonics, 2020, p. 114332N.

[16] A. W. Harley, A. Uikes, and K. G. Derpanis, “Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval,” in Proceedings of International Conference on Document Analysis and Recognition, 2015, pp. 991–995.

[17] A. E. Alshehri and A. Alshehri, “The Supatrantique,” 2021.

[18] C. B. Rabah, G. Coatrieux, and R. Abdelfattah, “The Supatrantique,” 2021.
[19] A. Berenguel, O. R. Terrades, J. Lladós i, and C. Cañero, “Evaluation of Texture Descriptors for Validation of Counterfeit Documents,” in *Proceedings of IAPR International Conference on Document Analysis and Recognition*, vol. 1, 2017, pp. 1237–1242.

[20] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen, “Revisiting Training Strategies and Generalization Performance in Deep Metric Learning,” in *Proceedings of International Conference on Machine Learning*, 2020, pp. 8242–8252.

[21] P. Bateni, R. Goyal, V. Masani, F. Wood, and L. Sigal, “Improved Few-shot Visual Classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 93–14 92.

[22] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, “Deep Anomaly Detection for Generalized Face Anti-spoofing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[23] J. Hu, J. Lu, and Y.-P. Tan, “Discriminative Deep Metric Learning for Face Verification in the Wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1875–1882.

[24] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, “No Fuss Distance Metric Learning using Proxies,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 360–368.

[25] M. J. Allen, *Foundations of Forensic Document Analysis: Theory and Practice*. John Wiley & Sons, 2015.

[26] A. Ferreira, L. Bondi, L. Baroffio, P. Bestagini, J. Huang, J. A. Dos Santos, S. Tubaro, and A. Rocha, “Data-driven Feature Characterization Techniques for Laser Printer Attribution,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1860–1873, 2017.

[27] L. Yang, *Ink-Paper Interaction: A Study in Ink-jet Color Reproduction*. Linköping University, 2003.

[28] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A Unified Embedding for Face Recognition and Clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[29] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[31] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[33] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How Transferable are Features in Deep Neural Networks?” in *Proceedings of the International Conference on Neural Information Processing Systems-Volume 2*, 2014, pp. 3320–3328.

[35] B. Bayar and M. C. Stamm, “Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, 2018.

[36] L. Yang, *Ink-Paper Interaction: A Study in Ink-jet Color Reproduction*. Linköping University, 2003.

[37] T. Thongkanwitoon, H. Maammar, and P.-L. Dragotti, “An Image Recapture Detection Algorithm based on Learning Dictionaries of Edge Profiles,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 953–968, 2015.