scientific data

OPEN

DATA DESCRIPTOR

Synthetic dataset of ID and Travel Documents

Carlos Boned[1,4], Maxime Talarmain[1,4], Nabil Ghanmi[2], Guillaume Chiron[2], Sanket Biswas[1], Ahmad Montaser Awal[2] & Oriol Ramos Terrades[1,3] ✉

This paper presents a new synthetic dataset of ID and travel documents, called SIDTD. The SIDTD dataset is created to help training and evaluating forged ID documents detection systems. Such a dataset has become a necessity as ID documents contain personal information and a public dataset of real documents can not be released. Moreover, forged documents are scarce, compared to legit ones, and the way they are generated varies from one fraudster to another resulting in a class of high intra-variability. In this paper we introduce a dataset, synthetically generated, that simulates the most common, and easiest, forgeries to be made by common users of ID documents and travel documents. The creation of this dataset will help to document image analysis community to progress in the task of automatic ID document verification in online onboarding systems.

Background & Summary

The development of remote identity authentication systems, which include both biometrics and ID and travel documents verification, has increased and spread since the advent of the COVID-19 pandemic. These authentication systems have allowed people to work and to develop their business activities out of their offices as public administration, banks, and productive industries and many services have adopted them in their usual workflow. These services offer online enrollment, thus, avoiding the user's physical attendance by requiring a selfie and a picture of their ID document to authenticate them. However, cybercrime has taken advantage of society's vulnerabilities and evolve towards more sophisticated threats. As pointed out by the IOCTA 2020 report[1], "the fundamentals of cybercrime are firmly rooted, but that does not mean cybercrime stands still. Its evolution becomes apparent on closer inspection, in the ways seasoned cybercriminals refine their methods and make their artisanship accessible to others through crime as a service". In particular, fraudsters may take advantage of these vulnerabilities by forging ID documents to alter information or hide their real identity. Consequently, new developments on identity authentication systems must include advanced AI tools to reliably ensure citizen's security and protect the online services.

A key tool to ensure citizen's identity in a digital environment, among others, is the detection of forged ID and travel documents when they enroll to online services. This *Presentation Attack Detection* (PAD) tool must compare an image, or video, most likely acquired by mobile devices, of citizen's ID documents and to assess if such ID document images, or videos, corresponds to a *bona fide* document or not. Given the current legislation on data protection, as the GDPR in the EU, publishing real ID documents data is restricted to those in which citizens has provide explicit consent. Consequently, it is difficult to gather enough data to estimate model parameters to detect forged documents and sophisticated AI models that generate synthetic ID and travel Documents have been developed[2]. These models train Generative Adversarial Networks (GANs) to simulate ID Documents containing information from non-existing people. Despite being models that generate quite realistic ID Documents, the generated data is useless for PAD tasks as they do not contain the security features that documents of this kind usually have added them.

The current trend is therefore to detect ID Documents that have been altered by detecting unexpected changes on the document texture, text or identity photo location by means of basic image processing techniques. Thus, GAN-based models are proposed to generate ID document images from a limited set of templates under three typical *presentation attack instruments*[3] (PAI), namely: *composite*, *print* and *screen*, as they are defined by the International Standard (ISO/IEC 30107)[4] and performance evaluation of general purpose classification networks on two tasks: *composition detection* and *source detection* is done. The first task aims at detecting the

[1]Computer Vision Centre, Bellaterra, 08193, Spain. [2]IDNow, 122 Rue Robert Keller, 35220, Cesson-Sévigné, France. [3]Universitat Autònoma de Barcelona, Dep. Computer Science, Bellaterra, 08193, Spain. [4]These authors contributed equally: Carlos Boned, Maxime Talarmain. ✉e-mail: oriolrt@cvc.uab.cat

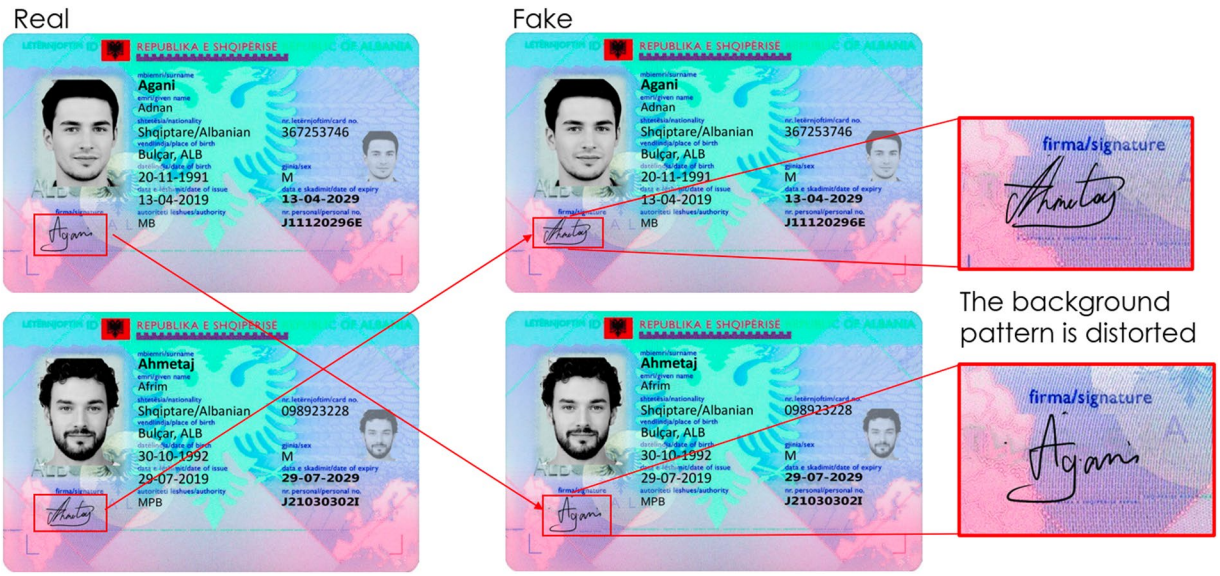

Fig. 1 Crop & Replace Composite PAI example. The signature of two ID documents of the same nationality is replaced.

composite PAI while the second task aims at detecting whether the ID Document comes from a *bona fide* document or a *printed* or a *screened* copy of a *bona fide* document. In any case, synthetic datasets used on that experiments are not published and cannot be actually used for benchmarking purposes[5].

The goal of this work is to release to the community a public dataset for PAD purposes given the above-mentioned PAI tasks. This dataset is an extension of the MIDV2020 dataset[2], which is the largest publicly available identity documents dataset with variable artificially generated data. The whole generation of forged data, as well as the algorithms developed for their generation, do not pose a high risk of malicious use. The entire process has been validated and approved by the Committee on Ethics in Research from the Autonomous University of Barcelona. In summary, the main contributions are:

- The proposed dataset, the SIDTD dataset, contains the original MIDV2020 images and videos, which will compose the corpus of *bona fide* documents, together with a set of images and videos, that are the altered version of the MIDV2020 ID Documents and compose the corpus of forged documents.
- Together with the SIDTD dataset we also publish the code repository used to generate forged documents. The implemented PAIs reproduce basic image editing operations to provide training and evaluation data to ID Document verification systems.
- A set of pre-defined data partitions for the following model performance evaluation: hold-out, k-fold cross validation and few-shot evaluation. Few-shot evaluation will allow evaluate progress in ID Document verification task on the proposed data set in a more real, and challenging, scenario.

Methods

As explained above, the SIDTD dataset is an extension of the MIDV2020 dataset[2] (http://l3i-share.univ-lr.fr/MIDV2020/midv2020.html). Initially, the MIDV2020 dataset is composed of forged ID documents, as all documents are generated by means of AI techniques. These generated documents are considered in the SIDTD dataset as representative of *bona fide*. On the other hand, the documents generated as described in this section will be considered as being *forged* versions of them. The corpus of the dataset is composed by ten European nationalities that are equally represented: Albanian, Azerbaijani, Estonian, Finnish, Greek, Lithuanian, Russian, Serbian, Slovakian, and Spanish.

Forged ID Document images generation. We employ two techniques for generating *composite* PAIs: *Crop & Replace* and *inpainting*. The *Crop & Replace* technique is a fundamental image processing approach that involves the exchange of information between two identification (ID) documents of the same class. This is achieved by cropping a specific region from one ID document and replacing it with corresponding information from another ID document, as illustrated in Fig. 1. To mitigate the risk of creating a perfect match and ensure the artificial documents are indistinguishable from their authentic counterparts, we introduce a *shift* parameter. This parameter determines the offset for the exchanged regions. Thus, we define a range $[-n, n]\backslash\{0\}$, where $n \in \mathbb{N}$, for the random setting of shift parameters along both the x-axis and y-axis. The shift value 0 is excluded to prevent perfect matching. The introduced shift parameter induces a border effect resulting from texture discontinuity, which must be detected by the PAD method.

Fig. 2 Inpainting Composite PAI example. The name field in the ID document is replaced by the same content but changing the font. Real ID Document (left) and Fake ID Document (right).

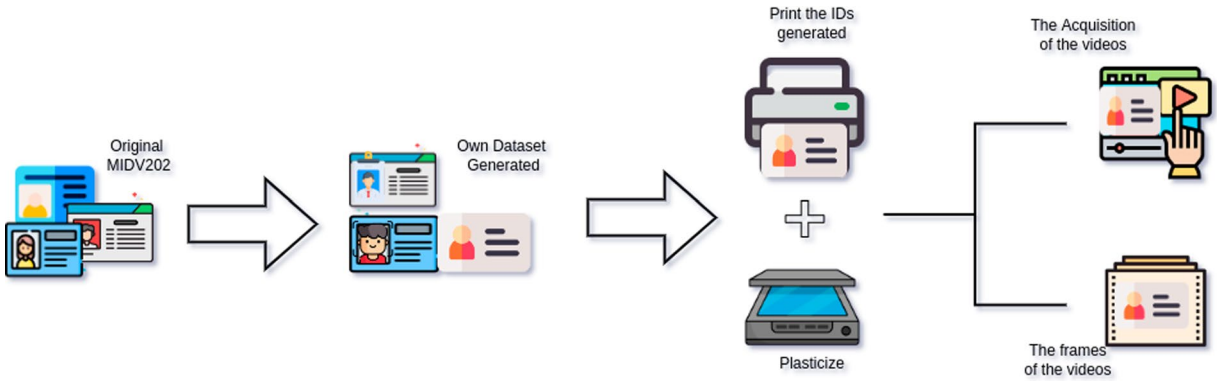

Fig. 3 Pathflow followed to generate forged videos and clips.

Conversely, the *Inpainting*[6] technique is a sophisticated image processing method that involves replacing a small region of an image while maintaining a realistic look and feel. This technique is commonly applied in post-production to remove people from pictures or architectural artifacts from images and movies. In the context of ID documents, *Inpainting* can be applied to eliminate personal information, such as names or surnames, from textual fields, replacing it with false information of the same nature.

First, *Inpainting* is employed to remove the original information by generating a realistic background that covers the text. Then, the size of the newly added text is computed through interpolation and inference based on the information surrounding the text fields, and the font is randomly selected from a set of available font types. An illustrative example of a forged ID document is presented in Fig. 2, where the name was regenerated using the *Inpainting* technique.

Depending on the shift value in the *Crop & Replace* technique and the chosen fonts and text sizes for each manipulation, the appearance of the generated ID documents can range from easily distinguishable to humans to extremely subtle and indistinguishable alterations.

Forged ID Document videos generation. As the original MIDV2020 dataset contains videos, and clips, of captured ID Documents with different backgrounds, we add the same type of data for the forged ID Document images generated using the techniques described in the previous section. The protocol employed to generate the dataset is as follows: We printed 191 counterfeit ID documents, created using the tools detailed in the previous section, on paper using an HP Color LaserJet E65050 printer. Then, the documents were laminated with 100-micron-thick laminating pouches to enhance realism and manually cropped, as depicted in Fig. 3.

CVC's employees were requested to use their smartphones for recording videos of forged ID documents from SIDTD. This approach aimed to capture a diverse range of video qualities, backgrounds, durations, and light intensities. The resulting dataset includes videos from various smartphone brands such as Samsung (Galaxy A5, Galaxy A52, Galaxy A53 5G, Galaxy A70, Galaxy M12, Galaxy M21, Galaxy M31s, Galaxy S10e, Galaxy S20+ 5G, Galaxy S21, Galaxy S22+), Xiaomi (Mi 10T Pro 5G, Mi 8, Mi 9 Lite, Mi A3, Mi Max 2, POCO M2, POCO X3 Pro, Redmi Note 7 pro, Redmi Note Pro 11+), OnePlus (OnePlus 7 Pro, OnePlus 6T), Apple (iPhone 13, iPhone 12, iPhone 11, iPhone 8), Motorola (Moto G12, Moto G31), Google (Pixel 4a), Oppo (Realme C2), each offering a broad spectrum of camera properties (see Fig. 4b for the distribution of camera resolutions in megapixels).

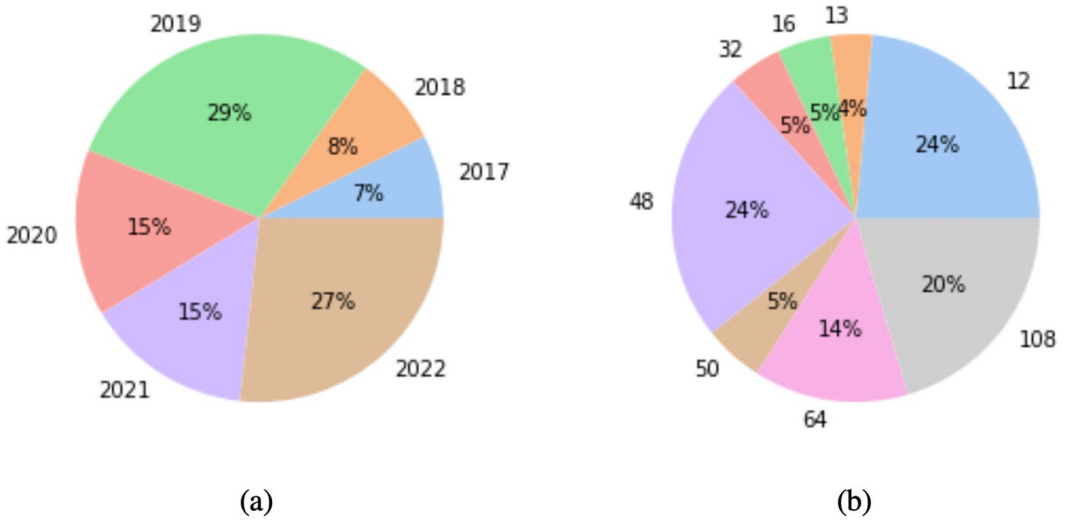

Fig. 4 Distribution of videos according to: (**a**) phone year of release (**b**) resolution of the smartphone main rear camera (in megapixels).

Fig. 5 Examples of SIDTD video clips of fake documents with different backgrounds, lightening and devices: (**a**) Finnish ID Document captured with natural light with table background recorded with Xiaomi Mi Max 2 (**b**) Spanish ID Document captured with natural light with outside floor background recorded with Samsung Galaxy A70 (**c**) Albanian ID Document captured with low lighting with chair background recorded with Xiaomi Redmi Note Pro 11+ (**d**) Russian passport captured with artificial indoor light with keyboard background recorded with Xiaomi Mi A3.

The recorded videos have relatively short durations, ranging between 4s and 13s, with an average of 7s. This duration is similar to *bona fide* ID document videos, which also average around 7s, varying between 4s and 12s. Overall, this procedure not only ensures diversity in the dataset but also enriches it with a variety of camera properties and conditions for robust model training. Most of the videos (approximately 85%) were captured using smartphones released within the last four years, as depicted in Fig. 4a. Also, image quality does not only depend on the resolution of the smartphone's primary rear camera[7] but several other parameters, such as image enhancement (improving brightness, contrast, colors, and reducing noise) and sensors features (including the number of sensors, sensor size, and sensor quality) play crucial roles in determining overall image quality. As these parameters vary from one smartphone to another, we cannot directly infer the image quality from the information showed in the Fig. 4b, we can thus note that the images in our dataset have been captured with a wide variety of devices using different image enhancement methods. Despite we gave the same instructions to each person, the angles, the movement, the video duration, the position of the document vary from one person to another. It causes more variability to the dataset and it results in a high diversity that will help to reduce model overfitting, see Fig. 5.

Finally, we extracted video clips from the recorded videos, every 6 frames as it was done for the MIDV2020 dataset. We annotated each corner of the identity document automatically using SmartDoc 2017's video capture method[8] based on the open source code they published on Github (https://github.com/smartdoc2017-competition/dataset_creation). The annotations are provided in JSON format with the same annotation structure as the one made for the MIDV2020 dataset.

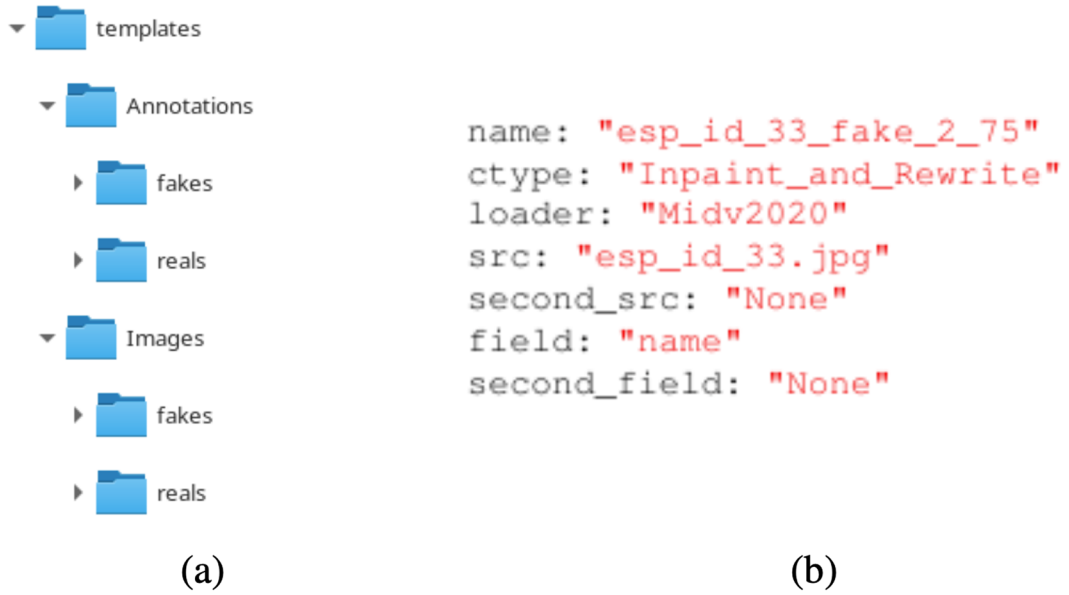

Fig. 6 (**a**) Example of folder structure corresponding to template document data after downloading it. (**b**) Metadata example of a forged document.

Private Dataset. To compare the SIDTD dataset to real datasets, we validate it with data from the company *IDNow* (https://www.idnow.io) as partners in the European project funding this work. This private dataset This private dataset was gathered by IDnow. Only authorized employees from IDnow are the ones granted to access to it. This dataset from IDnow is composed of real-life ID Document images, captured using various devices (scan, smartphone) without any constraint, and extracted from the flow passing through the IDCheck solution of IDNOW. Due to data privacy constraints, this data set cannot be shared with anyone outside the company, and thus could not be used in any other similar studies by other researchers. As in a real-world dataset, there are much less forged ID documents than *bona fide* ones, the sub-set of forged ID documents is mainly created by the following *Composite* PAI techniques, which are similar to the *Crop & Replace* technique described above:

- **Copy & Paste** operations inside the same document. The personal data fields from a real document are firstly located. Then, a source and a target fields are randomly selected and the content of the source field is copied and pasted in the target field.
- **Copy & Move** operations. A forged document is created from a real document by replacing its original identity photo with another one randomly selected from a given set of identity photos.

This private dataset is finally composed of 1, 900 *bona fide* examples and their corresponding 1, 900 *forged* examples. All the documents of this data set, whether *bona fide* or forged, are of various types (identity car, passports, driving license, resident permit, etc.) from different countries (France, Spain, Italy, Romania, etc.).

Data Records

The dataset is available at the Research data Repository (CORA) repository[9] and the TC-11 repository[10]. The SIDTD dataset contains images of ID documents of three different types: templates, videos, and clips. Each type includes both *bona fide* and *forged* documents. Data records foll all three types follows the same folder structure, see Fig. 6a for images of type templates. Data is split into Annotations and Images folders. Then, each of them is also divided in *bona fide* (real) and *forged* (fakes) data. The Images folder contains all the image documents in jpg format. Video files are saved in mp4 format. The Annotations folder also follows the same structure than the Images folder but containing metadata files. The contents of the *fakes* and *reals* folders in the Annotations folder slightly varies depending on the kind of data. For *real* data, as they belong the MIDV2020 instances, we keep the JSON files structure to preserve MIDV2020 dataset consistency in the SIDTD dataset. For *fake* data, we generated a JSON file with information related to the document generation process, the kind of PAI applied to the ID Document and other relevant information to reproduce, if needed, the *forged* image, see Fig. 6b. In Table 1, we describe the main metadata stored in the JSON files. The field *name* contains the name of the image file, while the field *ctype* contains the PAI technique applied to generate the forged document. Fields *src* and *field*, refers to the original image in the MIDV2020 dataset and the modified field. In case, the PAI technique requires a second image, and field, they are reported in the fields *second_src* and *second_field*, respectively.

The number of generated metadata files are the same as the generated images. It consists of 1,222 forged templates and 191 videos of forged documents. Clips are video frames, sampled every 6 frames. In total, we have extracted 7,214 clips from fake documents and 68,409 from real documents, Table 2. Generated videos

Class	Type	Field name	Description	Required
Bona fide	Templates	filename	Name of the *bonafide* ID document image	✓
		regions	List of regions that contains *shape* and *region* attributes	✓
		shape_attributes	Dictionary with fields: *name*, *x*, *y*, *width* and *height*	✓
		region_attributes	Dictionary with field: *field_name* and *value*	✓
Forged	Templates	name	Name of the ID Document image	✓
		ctype	Type of the applied PAI	✓
		loader	Name of the dataset used as *Bona fide* data	✓
		src	Name of the *Bona fide* ID document image	✓
		field	Name of the field modified in the *forged* document	✓
		second_src	Name of the secondary *Bona fide* ID document image used to generate *Crop* & *Replace* PAIs	✗
		second_field	Name of the secondary field used to *Replace* a field in the main ID Document	✗
Forged	Clips	filename	Name of the ID Document image	✓
		regions	List of regions that contains *shape* and *region* attributes	✓
		shape_attributes	Dictionary with fields: *name*, *all_points_x* and *all_points_y*	✓
		region_attributes	Dictionary with field: *field_name*	✓

Table 1. Description of main metadata of annotated document images. *Bona fide* instances keep MIDV2020 annotations.

		MIDV2020	Generated
source	Template	1,000	1,222
	Video	1,000	191
	Clip	68,409	7,214

Table 2. Statistics of the SIDTD dataset.

```
"00.jpg353386": {
  "filename": "00.jpg",
  "regions": [
    {
      "shape_attributes": {"name": "rect", "x": 397, "y": 211, "width": 86, "height": 29},
      "region_attributes": {"field_name": "name", "value": "Refik", "features": {"lowercase": true}}
    },
  etc],
}
```

Fig. 7 Example of an annotation from the MIDV2020 dataset.

metadata are refereed to the clips and they contain additional information about the position of the document image in each video frame. To keep consistency, we follow the same JSON file structure as it appears in the annotations from the clips of the original MIDV2020 dataset, see Fig. 7. Thus, the original MIDV2020 data, and metadata, are found, respectively, in the folder real of Images and Annotations folders. The field *filename* contains the name of the image file and the field *regions* is a list of text regions composed of rectangles. For each rectangle the bottom-right coordinate (*x*,*y*) is given, together with its width, height and text transcription.

We added the bounding box coordinates of the document location within the video clip to the JSON files. We follow the same bounding box representation used for the original MIDV2020 dataset, as shown in Fig. 8, to make data management as simple as possible. The field *filename* refers to the name of the clip and the field *regions* essentially contains the coordinates (at pixel level) of the modified region clockwise. Finally, we provide also the ID documents images from video clips, cropped and dewarped to remove image background and simplify data usage.

Data is partitioned to allow three model validation techniques: **hold-out, k-fold cross-validation** and **few-shot**. The code provided allow users to define any partition for these three model validation schemes. However, we also provide some pre-defined data partitions to make comparisons between models easier and fairer. Data is randomly sampled and it is, by default, split into 80%-10%-10% for the hold-out validation and split it into 10 folds for the k-fold cross-validation. For few-shot, 6 nationalities ID documents are randomly chosen for the meta-training and the remainder 4 for the meta-testing. However, for fair comparison between models we also provide a predefined partition in which training, validation and test instances are always the same for both, the hold-out and the k-fold cross-validation.

Pre-defined partitions are given by CSV files, as described in Table 3. The information of these CSV files basically consists of relative path to images, their label (*bona fide* or *forged*) and their class (country of issue of the

```
{
    "filename": str,
    "regions": [{"shape_attributes": {
                    "name": "polygon",
                    "all_points_x": [682,693,142,138],
                    "all_points_y": [188,1103,1068,237]
                },
                "region_attributes": {
                    "field_name": "doc_quad"}
                    }
}
```

Fig. 8 Example of the annotation format for the clips extracted from the videos.

Fields	Description
label_name	One of the two following values: *reals* and *fakes*
label	0 for *reals* instances and 1 for *fakes*
image_path	Relative path to the image
class_name	Abbreviation of the nationality of the ID Document
class	Integer number to denote the nationality of the ID Document
metaclass	One of the two following values: *ID* and *passport* (only for few-shot experiments)

Table 3. Fields of the CSV files generated for data partition.

Filenames	Validation technique	Class Distribution	Type of image
hold_out_split	Hold-out	Balanced	Templates
kfold_split	K-fold cross-validation	Balanced	Templates
few_shot_split	Few-shot	Balanced	Templates
unbalanced_hold_out_split	Hold-out	Unbalanced	Clip-cropped
unbalanced_kfold_clip_split	K-fold cross-validation	Unbalanced	Clip
unbalanced_kfold_split	K-fold cross-validation	Unbalanced	Clip-cropped
unbalanced_few_shot_split	Few-shot	Unbalanced	Clip-cropped

Table 4. Data partitions splits depending on the validation technique and whether the number of *bona fide* and *forged* classes are balanced, or not.

ID Document). For each validation technique we have 2 pre-defined data partitions: balanced and unbalanced. Unbalanced partitions use the clips extracted from the 191 videos recorded from the generated *forged* data, see Table 4. Conversely, balanced partitions uses the full set of *forged* data but using Templates instances. Both partitions: balanced and unbalanced, are consistent as samples in training, test and validations are the same in both kinds. So, if in the future we increase the recorded videos of *forged* data to all data, the splits of balanced and unbalanced data will be same but using Templates or Clip-cropped data.

Technical Validation

As described in section Methods, *forged* data has been generated by employing two composite PAI from the already annotated *bona fide* data. The parameters used to automatically generate *forged* data have manually set after visual inspection of few samples. As the goal of this dataset is to provide data that help automatic systems to be trained to detect suspicious ID, and travel, documents, the generated *forged* data must relatively easy to be spotted by any person after visual inspection. The generated *forged* data satisfies this requirement (type Template). We further needed to generate video instances of *forged* data in unconstrained scenarios. The main restriction was that ID documents were not cut in video recordings so they must appear complete as much as possible. This restriction was needed not only to ensure that document information was not missed but also to ensure document image detection, and dewarping. The quality of video recordings and the later clip cropping process has been ensure after visual inspection.

Document Verification tasks. To evaluate the utility of this dataset comparing real data, but not available in practice, in ID document verification task we have tackled performance evaluation of 5 representative deep learning models in the two subtasks described at the beginning of this paper: *Composite Detection* and *Source Detection*. For the *Composite Detection* task we have used the K-fold cross-validation partition on Template instances described in Table 4. For the *Source Detection* task we have used the K-fold cross-validation and the

Tasks	Composite Detection (K-fold, Balanced)		Source Detection					
			(K-fold, Unbalanced)		(Few-shot, Unbalanced)		Priv. dataset: K-fold, Balanced	
Dataset	**ACC**	**ROC AUC**	**ACC**	**ROC AUC**	**ACC**	**ROC AUC**	**ACC**	**ROC AUC**
EfficientNet	99.4 ± 1.0	1.00 ± 0.00	99.9 ± 0.1	1.00 ± 0.00	82.0 ± 2.6	0.899 ± 0.025	77.6 ± 3.4	0.865 ± 0.034
ResNet	98.1 ± 1.2	0.99 ± 0.01	99.9 ± 0.2	1.00 ± 0.00	90.6 ± 2.6	0.963 ± 0.017	82.5 ± 1.9	0.893 ± 0.014
ViT	55.2 ± 2.3	0.50 ± 0.04	97.5 ± 2.0	0.99 ± 0.02	91.0 ± 2.1	0.969 ± 0.011	52.7 ± 2.7	0.544 ± 0.032
TransFG	96.6 ± 1.5	0.99 ± 0.01	99.9 ± 0.2	1.00 ± 0.00	83.6 ± 3.3	0.907 ± 0.030	63.7 ± 3.1	0.741 ± 0.018
CoAARN	98.6 ± 1.6	0.99 ± 0.01	99.2 ± 1.2	0.99 ± 0.01	62.8 ± 5.5	0.675 ± 0.068	75.9 ± 2.9	0.826 ± 0.025

Table 5. Performance for the *Composite Detection* and the *Source Detection* tasks for the proposed dataset. Performance is also reported for the Private dataset.

Model	*Bona Fide*	*Inpainting*	*Crop & Replace*
EfficientNet	99.9 ± 0.10	99.4 ± 0.78	100.0 ± 0.00
ResNet	99.9 ± 0.18	99.5 ± 0.66	100.0 ± 0.00
VIT	99.3 ± 0.73	86.2 ± 8.06	93.8 ± 9.65
TransFG	99.9 ± 0.15	99.4 ± 1.64	99.6 ± 0.61
CoAARC	99.1 ± 1.15	98.3 ± 2.89	100.0 ± 0.00

Table 6. Accuracy of the reference deep learning models for each PAIs technique and *bona fide* samples on the k-fold cross-validation and unbalanced data partition.

Few-shot partitions on the Clip-cropped instances. Finally, to compare the SIDTD dataset on the later task to more real data we have evaluate the performance of the selected deep learning models on the private dataset: EfficientNet-B3[11], ResNet50[12], Vision Transformer Large Patch 16 (ViT-L/16)[13], TransFG[14] and Co-Attention Attentive Recurrent Network (CoAARC)[15,16]. The EfficientNet-B3 and ResNet-50 models are convolutional models widely used by the deep learning community for general purpose classification tasks. ViT-L/16 and TransFG are models inspired by the Transformer[17] encoder architecture from NLP models. The TransFG models is an extension of the ViT model. We use the same type of ViT model for TransFG as a base network: ViT-L/16.

The EfficientNet-B3, ResNet50 and ViT-L/16 are built-in models from PyTorch, which is a fully featured framework for building deep learning models, packages. ViT and CoAARC are trained with an input image resolution of 224 × 224 × 3, and EfficientNet-B3, ResNet50 and TransFG with image resolution of 299 × 299 × 3. Each model is pretrained on the ImageNet[18] dataset.

Table 5 show the obtained results for the different data partition and deep learning models. The high accuracy (ACC) scores for the *composite detection* task for most of the selected models (above 96%), could seem that the proposed dataset is quite simple for the proposed task. Similar conclusions could be drawn for the *Source Detection* task for the K-fold cross-validation partition on the Clip-cropped instances if we additionally compare both metrics (accuracy and the Area under the ROC-curve, ROC AUC) to the reported results on the private dataset. However, comparing to other similar datasets[5] the reported results show the proposed dataset is still challenging for the *composite detection* task. Table 6 report the accuracy for the *bona fide* and *forged* samples, depending on the PAI technique used to generate forged data. We can observe that overall the performance is similar for each subset of data.

Few-shot validation provides a better view about the current state of the art on ID Document verification techniques, as systems cannot be trained in practice with samples of existing ID Documents, passports, driving licenses, etc. worldwide. As shown in Table 5, the accuracy and ROC AUC scores for all reference models decrease significantly comparing to k-fold cross-validation data partition. The reported accuracy scores in Table 7 for *bona fide* data and *Inpainting* and *Crop & Replace* PAIs techniques show the utility of the proposed dataset to progress in the *source detection* task.

Usage Notes

The main functionalities of the dataset are divided into two sections: (i) Loading the dataset and (ii) generating a new dataset. All the necessary steps to use the dataset and generate new samples are described in the code repository. The dataloader is designed to download the full dataset with different partitions, loading them into the system memory. We can see some examples about how using it from Python or Bash shell scripting in Fig. 9. Moreover, we provide the functionality to generate more images using the techniques described in the Methods section. We show an example about how to generate new data based on the MIDV2020 dataset after installing the Python package in Fig. 10. Within the public Github repository, there is a subfolder located in the *data* directory, named *explore*. This folder contains more code examples showcasing the functions used for the creation of forged ID documents. Finally, the code is also ready to use the trained models and generate the csv files used to compute the reported results. We can see, in Fig. 11, an example reproducing the classification results with EfficientNet model (with and without GPU) on template instances of the ID documents.

Model	*Bona Fide*	*Inpainting*	*Crop & Replace*
EfficientNet	89.5 ± 3.07	86.8 ± 2.42	91.3 ± 5.04
ResNet	91.8 ± 2.67	91.2 ± 2.72	91.6 ± 8.51
VIT	91.6 ± 5.94	88.5 ± 4.80	96.1 ± 2.11
TransFG	85.7 ± 5.60	83.6 ± 6.50	87.1 ± 7.10
CoAARC	67.4 ± 4.60	66.2 ± 6.38	70.7 ± 8.27

Table 7. Accuracy of the reference deep learning models for each PAIs technique and *bona fide* samples on the few-shot and unbalanced data partition.

```python
from SIDTD.data.DataLoader.Datasets import *

data = SIDTD(download_original=False, custom_path_to_download=None).download_dataset("templates")
```

(a)

```
python data/DataLoader/Loader_Modules.py [--dataset DATASET] [--kind KIND] [--download_static]
    [--type_split TYPE_SPLIT] [--unbalanced] [-c|--cropped]
```

(b)

Fig. 9 Examples of code to load the dataset in (**a**) Python and (**b**) Bash shell.

```python
from  SIDTD.data.DataGenerator.Midv2020.Template_Generator import Template_Generator
from SIDTD.data.DataLoader.Datasets import *

## Downloadign our data just as an example
data = SIDTD(download_original=False).download_dataset("templates")

# get the abosulte path where the data is stored following the structure depicted above
path_dataset = "path to the downloaded dataset"

# generating the data
gen = Template_Generator.Template_Generator(absolute_path=path_dataset))

gen.create(sample=10)

gen.store_generated_dataset(path_store=None) #[None for dedault]
```

Fig. 10 Code example to generate new fake data.

```
# Test EfficientNet model with CUDA
python test.py --name='EfficientNet' --dataset='SIDTD' --model='efficientnet-b3'

# Test EfficientNet model with CPU
python test.py --name='EfficientNet' --dataset='SIDTD' --model='efficientnet-b3' --device='cpu'
```

Fig. 11 Example of code to load a pre-trained model, *EfficientNet* and test it onthe SIDT dartaset with, and without, GPU device.

Code availability

The code developed to download the data and prepare it to be used for models training and testing is available at the public code repository https://github.com/Oriolrt/SIDTD_Dataset. Every model is coded in Pytorch. All the scripts are coded with Python 3.7> and the setup.py is ready to install all the package dependencies. We strongly recommend to use Python environments to avoid package version dependencies issues. The models used to report results on the SIDTD dataset can be downloaded from the CVC repository.

Received: 4 March 2024; Accepted: 21 November 2024;
Published online: 18 December 2024

References

1. De Bolle, C. *et al*. Internet organised crime thread assesment (iocta). *EUROPOL* (2020).
2. Bulatov, K. B. *et al*. MIDV-2020: A comprehensive benchmark dataset for identity document analysis. *CoRR* **abs/2107.00396**, https://arxiv.org/abs/2107.00396 (2021).
3. Benalcazar, D., Tapia, J. E., Gonzalez, S. & Busch, C. Synthetic id card image generation for improving presentation attack detection. *IEEE Transactions on Information Forensics and Security* **18**, 1814–1824 (2023).
4. Information technology — Biometric presentation attack detection — Part 1: Framework. ISO Standard ISO/IEC 30107-1:2023, International Organization for Standardization (ISO), Geneva, Switzerland https://www.iso.org/standard/95925.html (2023).
5. Hamido, M., Mohialdin, A. & Atia, A. The use of background features, template synthesis and deep neural networks in document forgery detection. In *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 365–370, https://doi.org/10.1109/ICAIIC57133.2023.10067120 (2023).
6. Telea, A. C. An image inpainting technique based on the fast marching method. *J. Graphics, GPU, & Game Tools* **9**, 23–34 (2004).
7. Ovchar, I. Image quality is more than megapixels. *Petapixel* (2022).
8. Chazalon, J. *et al*. Smartdoc 2017 video capture: Mobile document acquisition in video mode. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, 11–16 (IEEE, 2017).
9. Boned, C. *et al*. Synthetic dataset of id and travel documents. *CORA.Repositori de Dades de Recerca*, https://doi.org/10.34810/data1815 (2024).
10. Boned, C. *et al*. Synthetic dataset of id and travel documents. *TC-11 Dataset repository,* https://tc11.cvc.uab.es/datasets/SIDTD_1/ (2024).
11. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114 (2019).
12. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
13. Dosovitskiy, A. *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
14. He, J. *et al*. TransFG: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 852–860 (2022).
15. Berenguel Centeno, A., Ramos Terrades, O., Lladós, J. & Cañero, C. Recurrent comparator with attention models to detect counterfeit documents. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, 1332–1337, https://doi.org/10.1109/ICDAR.2019.00215 (IEEE, 2019).
16. Wu, L. *et al*. Deep coattention-based comparator for relative representation learning in person re-identification. *IEEE transactions on neural networks and learning systems* **32**, 722–735 (2020).
17. Vaswani, A. *et al*. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
18. Deng, J. *et al*. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

Acknowledgements

SOTERIA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101018342. This content reflects only the authors' view. The European Agency is not responsible for any use that may be made of the information it contains. This work has been partially supported by the Spanish project PID2021-126808OB-I00, Ministerio de Ciencia e Innovación and the Departament de Recerca i Universitats of the Generalitat de Catalunya, DocAI reference 2021SGR01499.

Author contributions

O.R.T. conceived the experiments and the paper; C.B., M.T and S.B. implement the main Pyhton scripts, generated the data and conducted the experiments on the SIDTD dataset; N.G., G.C. and A.M.A., as IDnow employees, conducted the experiments on the private dataset. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to O.R.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.