FantasyID: A dataset for detecting digital manipulations of ID-documents

Pavel Korshunov, Amir Mohammadi, Vidit, Christophe Ecabert, and Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland
{pavel.korshunov, amir.mohammadi, vidit.vidit, christophe.ecabert, sebastien.marcel}@idiap.ch

Abstract

Advancements in image generation led to the availability of easy-to-use tools for malicious actors to create forged images. These tools pose a serious threat to the widespread Know Your Customer (KYC) applications, requiring robust systems for detection of the forged Identity Documents (IDs). To facilitate the development of the detection algorithms, in this paper, we propose a novel publicly available (including commercial use) dataset, FantasyID, which mimics real-world IDs but without tampering with legal documents and, compared to previous public datasets, it does not contain generated faces or specimen watermarks. FantasyID contains ID cards with diverse design styles, languages, and faces of real people. To simulate a realistic KYC scenario, the cards from FantasyID were printed and captured with three different devices, constituting the bonafide class. We have emulated digital forgery/injection attacks that could be performed by a malicious actor to tamper the IDs using the existing generative tools. The current state-of-the-art forgery detection algorithms, such as TruFor, MMFusion, UniFD, and FatFormer, are challenged by FantasyID dataset. It especially evident, in the evaluation conditions close to practical, with the operational threshold set on validation set so that false positive rate is at 10%, leading to false negative rates close to 50% across the board on the test set. The evaluation experiments demonstrate that FantasyID dataset is complex enough to be used as an evaluation benchmark for detection algorithms.

1. Introduction

Different financial services, like banks and insurance companies, use face and document verification systems to authenticate their users. This digital *Know Your Customer* (KYC) lets users take pictures of their ID (passports, residence permits, and driving licenses) using phone camera. The captured image is then typically compared to a selfie of the user to validate the authenticity of the document and the user's information. This process results in quick onboarding and increases the efficiency of the overall system.

(a) Chinese language

(b) Turkish language

Fig 1. **FantasyID:** Examples of original digital versions of FantasyID cards. The faces are of real people but the other biometric information is not real. The cards contain Guilloche patterns and design elements inspired by the ones used in official ID documents. We design 13 different card templates which are then physically printed with biometric details and recaptured to create 1086 *bonafide* images.

However, such KYC process presents a security risk, as one can use fake or forged documents to create a fraudulent account. The malicious user can potentially bypass the camera capture API and directly *inject* a forged image for authentication. Alternatively, an attacker can print the

(a) Persian language (b) French language (c) Arabic language (d) Russian language

Fig 2. Other examples of original digital versions of FantasyID cards.

digitally forged document, they have access to ID documents printing equipment, and capture it using a 'normal' onboarding process thus fooling the system. This vulnerability of KYC systems is exacerbated by the availability of rapidly improving image generation and editing methods [14, 6, 8, 22] allowing realistic-looking forged documents to be generated within minutes. To mitigate these risks, KYC process requires appropriate detection algorithms to flag forged documents.

There is a lack of publicly available datasets suitable for the detection of forged ID documents. The main reason is the restriction by official authorities on tampering of legal documents and publicizing sensitive personal information. Previous public ID datasets [2, 5, 4, 3, 1, 20] are either not meant for the manipulation-detection task, lack diversity, or contain bonafide samples that are tampered versions of official specimen ID documents (the watermark is manually or automatically removed).

To mitigate these issues, we introduce the first publicly available for commercial and non commercial use dataset, FantasyID[1], that contains both genuine and forged ID cards. Our ID cards (Fig. 1) are designed to resemble official ID documents (such as passports or ID cards) while avoiding legal problems associated with official ID document tampering, hence the name *Fantasy ID*. We created 13 templates representing genuine/bonafide IDs in the unique style of the following languages: Arabic, Chinese, Hindi, French, Persian, Portuguese, Russian, Turkish, Ukrainian, and English. All designs were crafted using Creative Commons 4.0-licensed source materials. The ID cards do not strictly follow the ICAO [12] standard for travel documents, because we used face images with public access licenses, so they do not follow strict passport-photo standards and because these cards are not meant to serve as official documents. However, our ID cards contain the main elements of an ID document, such a facial image, different text, including official and personal information, design elements resembling some of the real-world documents and Guilloche patterns often present in the ID documents.

FantasyID dataset has the following unique and novel characteristics:

1. FantasyID is the first public dataset where the *bonafide* cards are not modified versions of some official ID cards (e.g., with a digitally removed word 'specimen'). We provide pristine *bonafide* cards, which is important, since tampered images will bias digital manipulations detection algorithms.

2. FantasyID contains IDs for several non-English languages and is created using design patterns that mimic the styles used in the corresponding cultures. Hence, the FantasyID dataset facilitates research in multilingual text manipulation detection.

3. We use the faces of real people to avoid biasing our *bonafide* to generated fake faces.

4. In addition to *bonafide* cards, we provide digital manipulations created using face swapping approaches, such as InSwapper and Facedancer [23], and text inpainting techniques, such as Textdiffuser2 [6] and DiffSTE [14] (see more details in Sec. 3.2).

In the following sections, we discuss the existing datasets of ID documents, the generative methods used to create forged documents, and the existing forgery detection approaches; describe the process of building the FantasyID dataset; and present evaluation results of baseline forgery detection algorithms to demonstrate how challenging and useful FantasyID is for forgery detection research.

2. Related Work

We summarize different ID datasets (Tab. 1) highlighting their intended purposes and evaluating their suitability for manipulation detection tasks.

MIDV-500 [2]. This dataset is based on 50 specimen copies of the real IDs sourced from Wikimedia Commons[2], with the term "specimen" digitally removed. The images were printed and then the video was captured using different devices and backgrounds. The dataset was created to recognize and analyze the IDs using mobile devices. The tampering of the originals makes them biased and they cannot be considered as *bonafide*.

[1] https://www.idiap.ch/paper/fantasyid

[2] https://commons.wikimedia.org/wiki/Main_Page

(a) Indian, iPhone 15 Pro (b) Ukrainian, Huawei Mate 30

Fig 3. **Printed Bonafide:** Examples of bonafide ID cards printed and captured with different devices.

MIDV-2019 [5]. It extends MIDV-500 by capturing the physical cards under large perspective distortion and low-light conditions. The primary use case of this dataset is similar to MIDV-500 for ID recognition and analysis. It also has the same tampering bias in the *bonafide* images.

MIDV-2020 [4]. This dataset digitally alters texts and faces of the 10 IDs in MIDV-500 to create 1000 images. The primary goal is to study the text and face region detection using this dataset. The authors propose to use these 1000 images as fake documents to evaluate forgery detection algorithms. However, the original MIDV-500 manipulated regions like the removed word "specimen" are not labeled. This creates a bias for the detection algorithms.

FMIDV [1]. It was proposed to detect guilloche pattern-based forgeries as they are a common security feature in official ID documents. This dataset is created by applying a copy-move attack on MIDV-2020 images but only in the non-text/face region. These kinds of forgery are interesting but their scope is limited as they do not cover biometrics like faces and texts. Additionally, they treat MIDV-2020 images as *bonafide*, which already has digital manipulations.

KID34k [20]. This dataset was specifically created for online identity card fraud detection. It contains images of digitally created 82 ID cards based on Korean driver's licenses and registration cards. Here, the biometric information used is fake for both text and faces. The main purpose of the dataset is to study *screen* and *paper* attack detection. Around 34662 images are created by recapturing images displayed on screens and printed using 2 kinds of paper printers. The dataset is limited in the diversity of style and language (*only* Korean) used, which is needed for the robust study of detection algorithms. Moreover, fake faces used to create *bonafide* leads to bias in the dataset.

BID [9]. BID dataset was proposed for ID analysis tasks, such as text and image region segmentation, optical character recognition (OCR), and ID recognition. To create the dataset, biometrics features such as personal details and faces were erased and fake details were added digitally to the original Brazilian ID cards. These changes lead to a corrupted *bonafide*, rendering the dataset not suitable for manipulation detection tasks.

SIDTD [3]. Another extension of MIDV-2020 dataset. The images from MIDV-2020 are considered *bonafide* and are used to generate *copy-replace* and *inpainting* based forgeries. These images are physically printed to create a *presentation* attack scenario. As mentioned above, employing MIDV-2020 images as *bonafide* is already biasing the system towards manipulated *bonafide*.

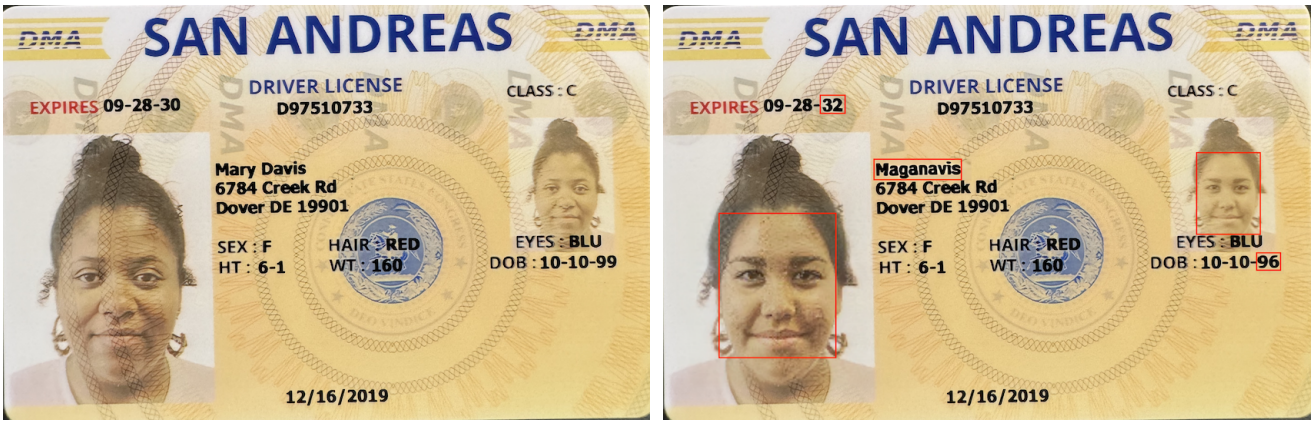
PAD [25] Recently proposed competition on PAD evaluated several methods with their private database of 300K ID cards. This database consists of *print*, *screen* and *composite* attacks. Since, we specifically study *injection* attack using digital image manipulation, comparison is out of scope with this work.

3. FantasyID Dataset

In this section, we provide an overview of the process of designing and building the *FantasyID* dataset, a diverse set of identity cards aimed at advancing research in forgery detection algorithms within biometrics. Our dataset consists of two main categories of ID cards: *bonafide*, which are the pristine cards, and *attacks*, which are different *digital* manipulation of pristine cards.

3.1. Creation of Bonafide Fantasy ID Cards

The first category of the dataset focuses on genuine *bonafide* fantasy ID cards. This process involved three key



(a) Bonafide English, captured with iPhone 15 Pro (b) Digitally manipulated

Fig 4. **Digital Forgery:** English language ID-card printed and captured with iPhone 15 Pro, and then manipulated by swapping face and altering text. Note that the box in red shows digitally manipulated areas.

Dataset	# IDs	Use Case	Source DB	Bonafide			Attack	
				Real Face	Pristine	Pr. Capt.	Digit. Manip.	Pr. Capt.
MIDV-500 [2]	50	ID Rec.	Web	✓	✗	✓	✗	✗
MIDV-2019 [5]	50	ID Rec.	MIDV-500	✓	✗	✓	✗	✗
MIDV-2020 [4]	1000	ID Rec.	MIDV-500	✗	✗	✓	✓	✗
FMIDV [1]	1000	Guilloche Det.	MIDV-2020	✗	✗	✓	✓	✗
KID34k [20]	82	PAD	Generated	✗	✗	✓	✗	✗
BID Dataset [9]	NA	ID Rec.	Real IDs	✗	✗	✗	✗	✗
SIDTD [3]	10	PAD	MIDV-2020	✗	✗	✓	✓	✓
FantasyID (Ours)	362	Man. Det.	Generated	✓	✓	✓	✓	✓

Tab 1. **ID Datasets.**. The table shows key characteristics of publicly available datasets for ID document analysis. Columns: **Real Face** indicates whether a real face was used on ID, **Pristine** means genuine bonafide cards were not manipulated in any way, **Pr. Capt.** indicates if the cards were printed and captured with a device's camera, and **Digit. Manip.** shows if cards were digitally manipulated to simulate a forgery attack. Only FMIDV was proposed for manipulation detection but restricted to manipulation in the guilloche patterns. Except for KID34k and BID, others are based upon MIDV-500 images and carry the bias of the manipulation in *bonafide* images. KID34k is limited in the diversity of style and language used on the cards. Furthermore, it uses generated faces for *bonafide* which will bias the detection algorithms. Our, FantasyID, provides pristine *bonafide* images with real faces along with diversity in card design and languages present. ID Rec. stands for ID Recognition and Man. Det. for Manipulation Detection. PAD is Presentation Attack Detection.

steps: ID card design and generation, printing, and image capture, respectively. We assume the scenario when a genuine user will capture their genuine physical card with a phone or scanner when enrolling into KYC service.

1. **ID Card Generation**: We generated 262 genuine ID cards for training set and additional 100 ID cards for test set, using random personal data (e.g., date of birth, names, city of origin) specific to each language (see Fig. 1 and Fig. 2 for some examples). These cards feature real faces from datasets such as the American Multiracial Face Database (AMFD) [7], Face London Research Dataset [10], High-Quality Wide Multi-Channel Attack (HQ-WMCA) [18] and a set of 100 face images with open license (public domain or CC-BY-4.0) manually downloaded by us from Flickr.

2. **Printing**: The generated cards were printed using Evolis Primacy 2 card printer[3] to emulate real-world ID cards. This step resulted in 362 high-quality printed ID cards *(600 DPI)*, prepared for further processing.

3. **Image Capture**: Digital images of each printed ID card were captured using three devices (Apple iPhone 15 Pro, Huawei Mate 30, and Kyocera TASKalfa 2554ci office scanner). See Fig. 3 for the examples of printed and captured *bonafide* cards. A set of the 362 printed cards was captured three times (once with each device) resulting in a total of 1086 *(362 × 3)* high-quality images of *bonafide* fantasy ID cards. We use 786 bonafide cards for train-val set and remaining 100 for testing.

[3] Evolis Primacy 2 Printer

Category	Sub-Category		Devices/Sources	Description and Purpose
Bonafide Cards	Generation	362	AMFD, Face London, WMCA, and Flickr datasets	Thirteen unique design styles in ten languages. Random but realistic personal info.
	Print	362	Evolis Primacy 2 printer	Printed on physical plastic cards.
	Capture	1086	iPhone 15 Pro, Huawei Mate 30, office scanner	Plastic cards were captured using three devices.
Forged Cards	Digital manipulation	786	InSwapper	Each face in captured IDs from *train-val* set is swapped with another face.
	Digital manipulation	786	Facedancer [23]	Each face in captured IDs from *train-val* set is swapped with another face.
	Digital manipulation (Attack-2)	150	Facedancer [23]	Faces in captured IDs from subset of *test* set are swapped.
	Digital manipulation	786	DiffSTE [14]	Parts of personal info in captured IDs from *train-val* set were replaced by another text.
	Digital manipulation	786	Textdiffuser2 [6]	Parts of personal info in captured IDs from *train-val* set were replaced by another text.
	Digital manipulation (Attack-1)	786	Finetuned-Textdiffuser2	Parts of personal info in captured IDs from *train-val* set were replaced by another text.
	Digital manipulation (Attack-3)	149	Finetuned-Textdiffuser2	Parts of personal info in captured IDs from *test* set were replaced by another text.

Tab 2. The summary of how different parts of FantasyID dataset were created.

3.2. Creation of Manipulated Fantasy ID Cards

The second category of the dataset deals with fake fantasy ID cards. These manipulated cards form the digitally altered attacks of FantasyID and represent the use case when an attacker is trying to subvert a KYC system by submitting a digital fake version of an ID card with face, name, or/and date of expiry altered (see an example in Fig. 4). The presentation of this fake card is done digitally by bypassing a sensor of the KYC system. This category is further divided into two sub-categories based on manipulation methods:

1. **Train-Val**: 786 bonafide cards which are captured using three different devices were digitally modified by swapping faces and inpainting text regions. We create two set of manipulations by using different swapping and inpainting methods, namely: (a) InSwapper (InsightFace[4]) for face and DiffSTE [14] for text (b) Facedancer [23] for face and Textdiffuser2 [6] for text. Fig. 4 demonstrates the results of these digital manipulations. Therefore, we have 786 bonafide and 1572 manipulated images in this set. We keep all 459 cards containing faces from HQ-WMCA dataset as val set and the rest of 1899 images as training set. The training set can be used to fine-tune a detection model.

2. **Test**: In order to create a challenging test set, we create distinct manipulations on the bonafide of train-val set, i.e., **Attack-1**. It consists of 786 images with *text-only* modification created by finetuning Textdiffuser2 [6]. The generated text regions are post-processed with the Segment Anything Model [15] to extract text regions with their background, and alpha blended with original background, where the text region is erased using LaMA-inpaint [24]. This approach creates text modifications that are hard to notice visually.

   Further to test the out-of-domain generalization, we create a bonafide set consisting of 300 images which are distinct from train-val set in terms of template design, faces, and textual details. We modify these 300 cards using different kind of manipulations: (b) **Attack-2** consists of *face-only* manipulation using Facedancer [23] on 150 ID cards and (c) **Attack-3** is created by changing *text-only* regions using the same approach as in Attack-1 but applying it to the 149 ID cards from the test set. In total, we have 300 bonafide and 1085 manipulated images in the test set.

Each manipulation category aims to reflect real-world attack scenarios for testing the robustness of detection algorithms. Tab. 2 summarizes the key steps, categories, devices, and purposes involved in the creation of genuine and fake FantasyID cards.

[4] https://insightface.ai/

4. Evaluation of Baselines on FantasyID

FantasyID dataset is the first dataset of ID documents where bonafide digital originals and the digital attacks on the face and text data are available in the public domain. Nevertheless, to demonstrate that this dataset also poses a challenge to algorithms designed to detect manipulations and document forgeries, we conducted a set of experiments evaluating baseline algorithms on this dataset. For the baseline algorithms, we have only considered methods that focus on generic synthetic image or local region manipulations detection, as oppose to more known methods of deepfake detection that typically focus on faces [13, 16].

In this paper, we used the following four state of the art algorithms for binary detection of manipulations in images: TruFor [11], MMFusion [26], UniFD [19], and FatFormer [17]. We have used the pretrained models provided by the authors. The fake images used in the training datasets of TruFor and MMFusion are tampered images where only some parts of the image are modified. Whereas the fake images seen during training of UniFD and FatFormer are GAN generated images where the full image is digitally generated. More details about the algorithms are given below.

- **TruFor [11]** uses a multi-branch Transformer encoder architecture to combine features from RGB images and Noiseprint++ images to predict an anomaly localization map, a confidence map, and a final score. Noiseprint++ [11] is a fully convolutional network trained to extract the subtle noise present in pristine images caused by imperfections in camera hardware or in-camera processing steps. In our binary detection experiments, we used the final score of TruFor.

- **MMFusion [26]** extends TruFor by adding more modalities to the encoder architecture. The authors propose to use Steganalysis Rich Model (SRM) filtered and Bayar convolution images in addition to the Noiseprint++ images. The integration of SRM and Bayar images is done in two ways: i) early fusion, when images (except for RGB) are passed into a convolution block and are merged into one image before being used as input to the encoder architecture, and ii) late fusion, when the encoder is repeated for each pair of RGB and one other image modality and their final features are combined. The weights of the RGB branches are shared in this case. In our binary detection experiments, we have used the early fusion variant of MMFusion because of its better performance.

- **UniFD [19]** is a simple approach that uses features from a pretrained frozen CLIP:ViT-L/14 [21] model and a linear classifier to detect fake images. The aim of UniFD is to detect fully synthetic images, so it may under-perform on images that are only partially modified.

- **FatFormer [17]** uses the CLIP:ViT-L/14 model similar to UniFD. However, it adds forgery-aware adapters to the ViT model that contains convolution and discrete wavelet transform operations. It also introduces language-guided alignment using the text encoder of CLIP to guide the image encoder to focus on forgery-related representations.

4.1. Evaluation Protocol and Metrics

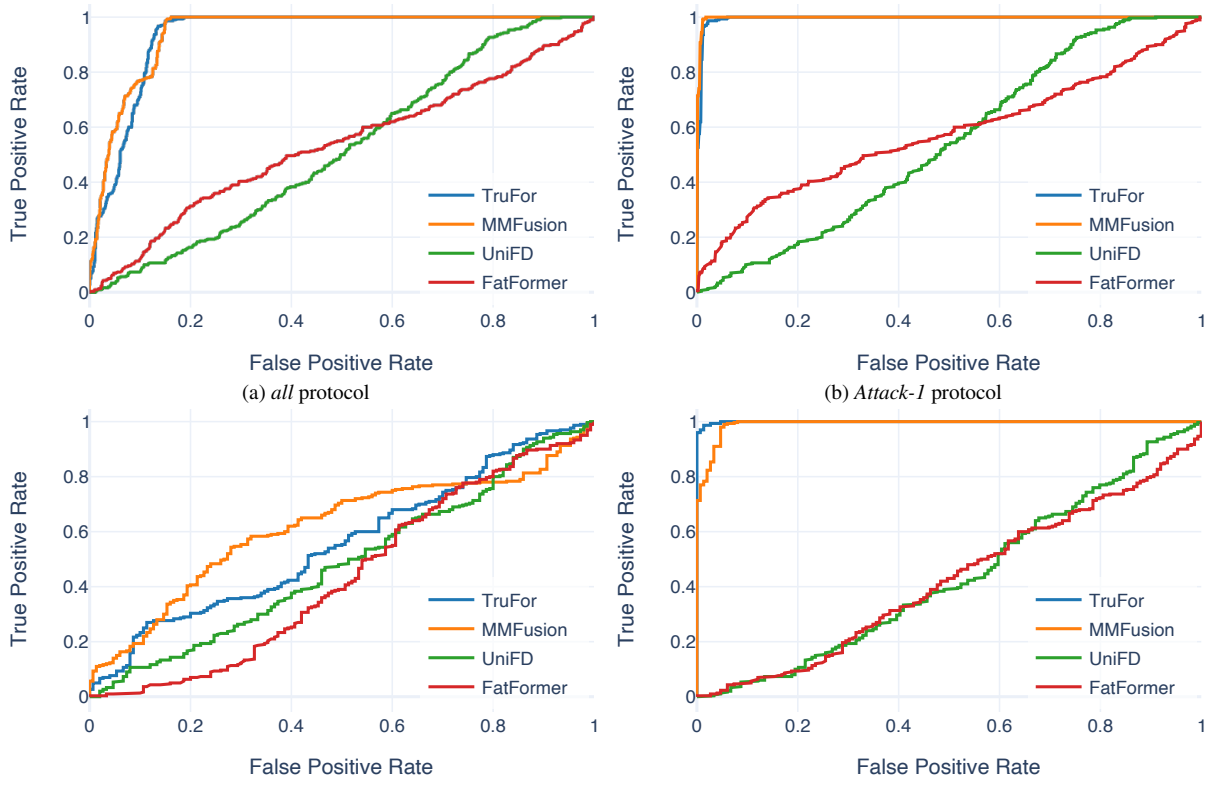
To make sure the performance of the baseline methods does not degrade due to the difference between training and testing, we applied the same preprocessing step to the input images that were proposed by the authors of the corresponding algorithms. TruFor and MMFusion baselines work on full resolution images. The input images to UniFD are resized so that the shortest side is 224 pixels and are then center cropped to obtain a 224×224 square image. The input images to FatFormer are first resized to 256×256 pixels and then center cropped to 224×224 pixels. We also evaluated both UniFD and FatFormer on zero-padded images to make the images square but that lead to worse performance.

For evaluation metrics, we report commonly used metrics for binary classification: false positive rate (FPR), where positives are *bonafide* images, false negative rate (FNR), and half total error rate (HTER), which is the average of FPR and FNR. These rates are computed on the *Test* set using a threshold estimated on the *Val* set at 'FPR$=10\%$'. In addition to these metrics, to be comparable with the metrics used by the authors of the baseline methods, we also report area under the curve (AUC) of ROC plots, balanced accuracy (ACC), and F1 score weighted by class on the *Test* set. ACC and F1 are computed using a fixed threshold of 0.5.

Even though we use the baseline algorithms *as is* without additional tuning on the training set of FantasyID, we evaluate them on the test set only. In this way, FantasyID test set can be used as a benchmark in the future to compare different methods for ID documents manipulation detection, while the train-val set can be used to tune the models. Hence, we compute metrics for the test set overall and for each of the three attacks (see Sec. 3.2 for details), to have a better understanding about the impact of each type of manipulation on the performance of the baselines.

4.2. Evaluation Results

We evaluated manipulation detection algorithms in terms of their binary detection performance. The evaluation is done separately for different digital manipulations, i.e, Attack-1, Attack-2, Attack-3, and *all* representing aggregated result of the three attacks.



(a) *all* protocol (b) *Attack-1* protocol

(c) *Attack-2* protocol (d) *Attack-3* protocol

Fig 5. ROC plots of the baseline detection methods on all samples and on each of the attacks from the test set of FantasyID.

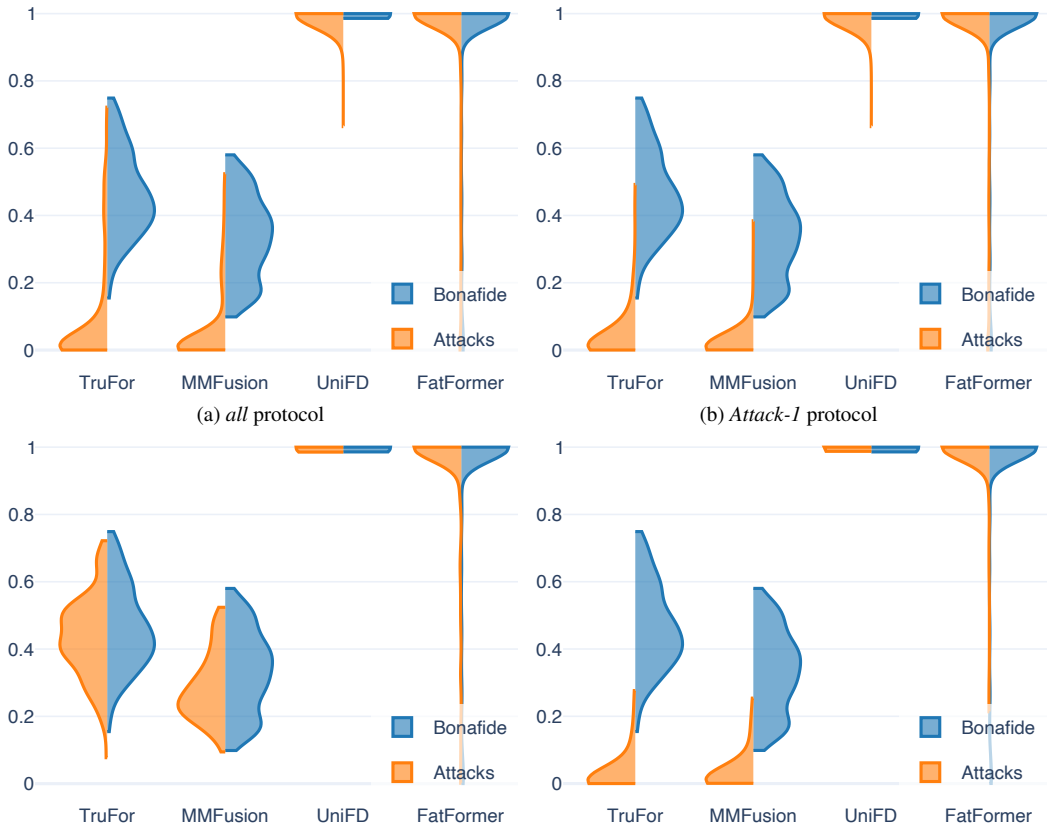
Model	Protocol	ACC	AUC	F1	FPR	FNR	HTER
TruFor	all	65.9	93.5	80.7	4.9	62.0	33.4
TruFor	Attack-1	67.8	99.5	79.0	0.1	62.0	31.1
TruFor	Attack-2	53.5	55.8	47.6	34.7	62.0	48.3
TruFor	Attack-3	67.8	99.9	55.3	0.0	62.0	31.0
MMFusion	all	55.1	94.4	73.7	4.0	47.7	25.8
MMFusion	Attack-1	55.2	99.8	67.0	0.1	47.7	23.9
MMFusion	Attack-2	54.5	61.5	29.8	28.0	47.7	37.8
MMFusion	Attack-3	55.2	99.1	30.0	0.0	47.7	23.8
UniFD	all	50.0	52.0	7.7	8.3	92.7	50.5
UniFD	Attack-1	50.0	54.3	12.0	7.4	92.7	50.0
UniFD	Attack-2	50.0	48.4	53.3	7.3	92.7	50.0
UniFD	Attack-3	50.0	43.5	53.5	14.1	92.7	53.4
FatFormer	all	48.8	53.5	15.6	6.5	92.3	49.4
FatFormer	Attack-1	49.3	57.6	20.4	1.1	92.3	46.7
FatFormer	Attack-2	48.3	43.6	53.8	24.0	92.3	58.2
FatFormer	Attack-3	46.7	42.3	51.8	16.8	92.3	54.6

Tab 3. Baselines methods evaluated on the test set of FantasyID, with digital forged ID cards as the negative set. ACC and F1 are computed using 0.5 as threshold while FPR, FNR, and HTER are computed using 10% FPR threshold from the validation set.

Tab. 3 shows the results of digital manipulation detection by our four baseline algorithms using metrics defined in Section 4.1. The results clearly demonstrate the advantage of TruFor [11] and its extension MMFusion [26], compared to CLIP-based FatFormer and UniFD methods. HTER and AUC metrics show that MMFusion is better at detecting several small manipulated text regions (Attack-1,3), with $HTER = 23.9\%$ and AUC above 99%. TruFor falls behind with $HTER = 31.1\%$ in the detection of manipulated text regions. Both FatFormer and UniFD show near random performance with average $HTER = 50\%$ on Attack=1,3. Their poorer performance can be attributed to the resizing operation (input image is 224×224) which attenuates the manipulation artifacts around small text regions.

While the performance is impressive for text manipulation detection, all the baselines fail to detect manipulation when only faces are edited, i.e, Attack-2. MMFusion performs better than random with $HTER = 37.8\%$, followed by TruFor, UniFD and Fatformer with $HTER = 48.3\%, 50.0\%, 58.2\%$, respectively. Attack-2, is created using Facedancer [23] which blends the facial regions with its background with a gaussian blur. Whereas, all the text manipulations are with simple alpha blending, hence creating a cut-paste attack. TruFor [11] is known to work well for these cut-paste manipulations, hence we observe higher performance on text manipulations and not on faces.

TruFor and MMFusion models treat forgery detection as a binary localization problem and have been trained to de-



(a) *all* protocol (b) *Attack-1* protocol

(c) *Attack-2* protocol (d) *Attack-3* protocol

Fig 6. Kernel density estimation (KDE) plots of scores from the baseline detection methods for bonafide and different attacks, including all attacks, Attack-1,3 (text manipulations) and Attack-2 (face manipulation).

tect fake images where only some parts have been tampered with, a scenario similar to our digital manipulation protocol. Whereas, UniFD and FatFormer were trained to detect fully generated images. This is reflected in both Tab. 3 and Fig. 5. ROC plots Fig. 5 show clear dominance of TruFor and its extension MMFusion.

Fig. 6 shows score distributions of different attacks and bonafide for all the baselines. These plots support Fig. 5 and clearly show Attack-2 (face manipulations only) as the most challenging to detect, since all the baselines have highly overlapping scores for bonafide and manipulated images. The scores for Attack-1,3 show clearer separation with MMFusion and TruFor whereas UniFD and FatFormer fail to distinguish between the bonafide and attacks. The plots illustrate the challenging nature of FantasyID with its diverse attacks.

Overall results demonstrate that even though such state-of-the-art algorithms as MMFusion and TruFor are able to detect the presence of local digital forgeries (mostly text) with a reasonable accuracy, their performance is far from practical applications standards. Since these attacks are easy to perform for a malicious actor, they continue to pose a serious threat to the detection systems.

5. Conclusion

In this paper, we presented the first publicly available with a permissive usage license dataset of fantasy ID documents that contain truly bonafide versions of the ID cards (in digital and printed/captured forms) and the fake cards with face and text manipulated. We tested the state-of-the-art forgery detection algorithms on both types of manipulations, demonstrating that binary detection of the forgeries are challenging for the current algorithms. An important future direction is the evaluation of detection algorithms localization performance, since often a forgery of an ID documents pertains only to small text regions such as a digit of an expiry date, making detection of such manipulations even more challenging.

We believe that this dataset, especially since it is publicly available also for commercial use, will help advance the research in forgery detection and localization in ID documents with the aim to protect users when they are onboarding in many popular online KYC services.

Acknowledgement

This work was funded by InnoSuisse 106.729 IP-ICT.

References

[1] M. Al-Ghadi, Z. Ming, P. Gomez-Krämer, J.-C. Burie, M. Coustaty, and N. Sidere. Guilloche detection for id authentication: A dataset and baselines. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, Poitiers, France, Sept. 2023. 2, 3, 4

[2] V. V. Arlazarov, K. B. Bulatov, T. S. Chernov, and V. L. Arlazarov. MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream. *Computer Optics*, 43(5):818–824, 2019. 2, 4

[3] C. Boned, M. Talarmain, N. Ghanmi, G. Chiron, S. Biswas, A. M. Awal, and O. R. Terrades. Synthetic dataset of ID and Travel Document, Jan. 2024. arXiv:2401.01858 [cs.CV]. 2, 3, 4

[4] K. Bulatov, E. Emelianova, D. Tropin, N. Skoryukina, Y. Chernyshova, A. Sheshkus, S. Usilin, Z. Ming, J.-C. Burie, M. M. Luqman, and V. V. Arlazarov. MIDV-2020: A comprehensive benchmark dataset for identity document analysis. *Computer Optics*, 46(2), Apr. 2022. 2, 3, 4

[5] K. Bulatov, D. Matalov, and V. V. Arlazarov. MIDV-2019: challenges of the modern mobile-based document OCR. In *International Conference on Machine Vision (ICMV)*, volume 11433, pages 717–722. SPIE, 2019. 2, 3, 4

[6] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision (ECCV)*, pages 386–402. Springer, 2024. 2, 5

[7] J. M. Chen, J. B. Norman, and Y. Nam. Broadening the stimulus set: introducing the american multiracial faces database. *Behavior Research Methods*, 53:371–389, 2021. 4

[8] X. Chen, B. Ni, Y. Liu, N. Liu, Z. Zeng, and H. Wang. SimSwap++: Towards Faster and High-Quality Identity Swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):576–592, Jan. 2024. 2

[9] A. de Sá Soares, R. B. das Neves Junior, and B. L. D. Bezerra. BID Dataset: a challenge dataset for document processing tasks. In *Anais Estendidos do XXXIII Conference on Graphics, Patterns and Images*, pages 143–146. SBC, 2020. 3, 4

[10] L. DeBruine and B. Jones. Face Research Lab London Set, 5 2017. 4

[11] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20606–20615, 2023. 6, 7

[12] I. C. A. O. (ICAO). Doc 9303: Machine readable travel documents, 2021. 2

[13] A. Jain, P. Korshunov, and S. Marcel. Improving generalization of deepfake detection by training for attribution. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2021. 6

[14] J. Ji, G. Zhang, Z. Wang, B. Hou, Z. Zhang, B. Price, and S. Chang. Improving diffusion models for scene text editing with dual encoders, Apr. 2023. arXiv:2304.05568 [cs]. 2, 5

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5

[16] P. Korshunov and S. Marcel. Improving generalization of deepfake detection with data farming and few-shot learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Jan. 2022. 6

[17] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10770–10780, 2024. 6

[18] Z. Mostaani, A. George, G. Heusch, D. Geissbuhler, and S. Marcel. The high-quality wide multi-channel attack (HQ-WMCA) database. *arXiv:2009.09703*, 2020. 4

[19] U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, 2023. 6

[20] E.-J. Park, S.-Y. Back, J. Kim, and S. S. Woo. KID34K: A dataset for online identity card fraud detection. In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 5381–5385, New York, NY, USA, Oct. 2023. Association for Computing Machinery. 2, 3, 4

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021. 6

[22] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund. FaceDancer: Pose- and occlusion-aware high fidelity face swapping. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3454–3463, 2023. 2

[23] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023. 2, 5, 7

[24] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 5

[25] J. E. Tapia, N. Damer, C. Busch, J. M. Espin, J. Barrachina, A. S. Rocamora, K. Ocvirk, L. Alessio, B. Batagelj, S. Patwardhan, R. Ramachandra, R. Mudgalgundurao, K. Raja, D. Schulz, and C. Aravena. First Competition on Presentation Attack Detection on ID Card, Aug. 2024. arXiv:2409.00372 [cs]. 3

[26] K. Triaridis and V. Mezaris. Exploring multi-modal fusion for image manipulation detection and localization. In *MultiMedia Modeling*, volume 14556, pages 198–211. Springer Nature Switzerland, 2024. 6, 7