Received: 14 March 2022 | Revised: 12 May 2022 | Accepted: 7 June 2022

IET Biometrics

DOI: 10.1049/bme2.12088

ORIGINAL RESEARCH

Pixel-wise supervision for presentation attack detection on identity document cards

Raghavendra Mudgalgundurao[1,2] | **Patrick Schuch**[1,2] | **Kiran Raja**[2] | **Raghavendra Ramachandra**[2] | **Naser Damer**[3]

[1]Nect GmbH, Hamburg, Germany

[2]NTNU - Gjøvik, Gjøvik, Norway

[3]Fraunhofer IGD, Darmstadt, Germany

Correspondence

Raghavendra Mudgalgundurao, Nect GmbH, Großer Burstah 21, Hamburg 20457, Germany.
Email: raghavem@stud.ntnu.no

[Corrections added on 14-July-2022, after first online publication: The placement of Figures 3, 4, 6, 7, 8, 9, 10 was modified while retaining their captions as is.]

Abstract

Identity documents (or IDs) play an important role in verifying the identity of a person with wide applications in banks, travel, video-identification services and border controls. Replay or photocopied ID cards can be misused to pass ID control in unsupervised scenarios if the liveness of a person is not checked. To detect such presentation attacks on ID card verification process when presented virtually is a critical step for the biometric systems to assure authenticity. In this paper, a pixel-wise supervision on DenseNet is proposed to detect presentation attacks of the printed and digitally replayed attacks. The authors motivate the approach to use pixel-wise supervision to leverage minute cues on various artefacts such as moiré patterns and artefacts left by the printers. The baseline benchmark is presented using different handcrafted and deep learning models on a newly constructed in-house database obtained from an operational system consisting of 886 users with 433 bona fide, 67 print and 366 display attacks. It is demonstrated that the proposed approach achieves better performance compared to handcrafted features and Deep Models with an Equal Error Rate of 2.22% and Bona fide Presentation Classification Error Rate (BPCER) of 1.83% and 1.67% at Attack Presentation Classification Error Rate of 5% and 10%.

1 | INTRODUCTION

ID documents such as passports and driver's licences that contain the face picture of the ID document owner are controlled and verified in many daily applications. Such documents are typically controlled and verified by humans or machines based on the application scenarios such as border crossing or purchasing age-restricted consumer goods. Furthermore, due to the increasing use of applications such as remote banking or ID control, especially in pandemic situations, ID cards are often controlled remotely [1, 2]. Following such motivation, a few studies have proposed approaches to verify the ID of a person using face image and the ID image in recent years [3, 4].

While remote ID card verification provides a convenience to the user, it allows users with malicious intent to misuse the copies (photocopied or digital) of ID cards for impersonating bona fide users. The digital copies of ID cards can be printed and presented to impersonate the ID of a different person. A more sophisticated attack can further present a digital copy of the ID card using various screens such as on tablets or smartphones. Thus, ID card verification and identification systems must detect print and replay attacks by accepting bona fide presentations and rejecting attack presentations. While the printed ID cards can be perceived as a low-quality attack, high-quality digital screens with higher resolution and higher visual quality can challenge the ID card verification systems to detect such attacks.

With the growing number of applications using remote ID card verification [5], very limited attention has been given to detect presentation attacks in ID card systems [6–8]. The necessity of a reliable ID card Presentation Attack Detection (PAD) system thus became inevitable. To address the vulnerabilities of presentation attacks in face recognition system, two general approaches were traditionally followed: (i) Hardware-based [9, 10] (ii) Software-based [11, 12]. Hardware-based

systems usually employ additional sensors to seek cues on presentation attack, for example, facial thermogram, specific reflection properties of the eye, perspiration, etc. Software-based systems utilise from the image/video to detect the attacks. Software-based feature approaches can be further classified into dynamic and static level approaches. Dynamic approaches include analysing challenge-response, blinks, specific movements of face while static approaches include learning classifiers on image descriptors like Histogram of Oriented Gradients [13], Local Binary Patterns [14, 15], Optical flow [16], and Histogram of Oriented Optical Flows [17].

Unlike in a face recognition system where a subject can be asked to respond to a dynamic challenge, ID cards cannot benefit from such approaches without involving a live face capture in the verification process of ID cards. Although the user can be asked to move the ID card while capturing, the attacker can follow such a procedure to bypass the PAD mechanism. Considering such challenges and not overloading the subject with additional face capture, we motivate our work to detect presentation attacks on ID cards directly without explicit user interaction. We, however, note a limited set of studies in this direction [7, 18] for tampering, print and replay attacks of ID cards. There are also studies based on reading the Near-Field Communication (NFC) chips that are embedded in the biometric International Civil Aviation Organization (ICAO) documents [19], but it should be duly noted that not all countries yet provide ICAO standard biometric documents to identify a person via NFC. In such cases, solely relying on the NFC feature for identification and verification would be rendered useless.

Motivated by the limited studies and the growing importance of attack detection in remote ID card verification, we present a new approach on ID card PAD in a remote verification scenario. The proposed approach is motivated by the observation that the appearance of bona fide presentations significantly differs from printed presentations and replay presentations, as noted in Figure 1. One can note the subtle differences in bona fide versus presentation attacks, which can be used for detecting presentation attacks. Based on such a hypothesis, we assert that pixel-wise supervision can effectively detect presentation attacks on ID cards. Inspired by the success of such binary supervision for attack detection in recent studies for face PAD [20] and morphing attack detection [21], we propose pixel-wise binary supervision for detecting attacks on ID cards. Specifically, in our proposed approach, we use a modified DenseNet121 [22] network with the complete image of the ID card crops resized to the network input requirements instead of using face crops.

As noted from a limited set of studies and unavailability of public datasets, we further create a new dataset of ID cards with both bona fide and attacks on two different kinds of ID cards from the European Union—the Personalausweis-ID card and the Aufenthaltstitel-Residence Permit [23]. Unlike the previous studies, the dataset is obtained from a real ID verification system, and the attacks are manually verified before marking them as attacks. The newly created dataset corresponds to in-the-wild settings where both bona fide and attack attempts are captured by various mobile phones. Further, the attacks in our new dataset correspond to images printed using different printers and a variety of displays used for replay attacks. The above mentioned database is in-house and due to General Data Protection Regulation, the dataset is not publicly available. Our contributions can be listed as below:

- A new database of ID cards is constructed in this work with 433 videos (10,677 images) along with 366 replay videos (79,926 images) and 67 print attacks (14,279 frames).

FIGURE 1 An illustration of Bona fide, Replay attack and Print Attack for ID cards. Bona fide images are clear and well defined with no moiré patterns. Replay attack images have different contrast, and images are blurry. The print images have different saturation due to the printer's quality. *Note - Personal information redacted to protect the privacy of the subjects in ID card

- We assume that the inputs to the trained model include only Personalausweis-ID card and the Aufenthaltstitel-Residence Permit.
- We present a new approach for ID card PAD using pixel-wise supervision for detecting presentation attacks and demonstrate the applicability of the proposed approach to detect presentation attacks efficiently.
- We present the benchmark performance evaluation using multiple handcrafted features and state-of-the-art deep learning models based on the newly created dataset. Specifically, we provide a benchmark against seven handcrafted approaches and three deep learning approaches.
- We further complement our results through a thorough analysis for explainability using multiple class activation maps (CAM). We provide an in-depth discussion on strengths and limitations of our proposed approach to facilitate the future works in this direction.

In the rest of the paper, we first present a set of related works in Section 2 and present the details of the newly created ID card attack database in Section 3. We then provide the detailed rationale and present our proposed approach in Section 4. We present a number of baseline evaluations and proposed approach in Section 5, and demonstrate the applicability of the proposed approach in detecting the attacks effectively. In Section 6, we present the heat maps of the networks to give an insight on the model learning the labelled images. In Section 7, we present results and experiments done towards the ablation studies. Section 8 analyses the cases for which the model prediction fails and measures that can be taken in future work to catch these scenarios. In Section 9, we explain the limitations of our current approach. Towards the end, we present a detailed analysis of the obtained results and provide some conclusive remarks along with the potential future works in Section 10.

2 | RELATED WORK

Presentation and other attacks have been studied for well-established modalities such as the face, iris and fingerprint for many years now [12, 24–26]. This section provides a set of related works for ID card verification.

Gonzalez et al. [7] recently proposed a two-stage approach for detecting presentation attacks on ID cards, specifically to detect print (from high-quality printers) and replay attacks (from digitally displayed on screens). The proposed two-stage network using MobileNet [27] is used to crop to the edges of the ID cards. A BasicNet is employed to classify bona fide, print and replay attacks created using Chilean national ID cards.

The BasicNet is trained with different inputs: i) Discrete Fourier Transform (ii) Steganalysis Rich Model (SRM) (iii) Error Level Analysis. Discrete Fourier Transform classifies the bona fide ID cards with 97.5% accuracy and a mean of 96.8% of attack ID cards. The drawback of this network is there is no widely available dataset to benchmark.

Zhou et al. [18] studied detecting tampering on ID cards. They proposed a method to detect tampering by providing more attention to artefacts rather than image content itself. The proposed approach consists of two R-CNN networks where the first network used the Red-Green-Blue (RGB) image and the second network used SRM of the RGB image (considered a noisy image) to detect the manipulated regions. With four standard image manipulation datasets such as National Institute of Standards and Technology Nimble 2016 [28], CASIA [29], COVER [30] and Columbia dataset [31], the authors demonstrated improved performance in both detecting and distinguishing between different tampering techniques.

Further, recently proposed DocFace [3] and DocFace+ [4] used an automatic system for matching ID document photos to live face images in real time. The proposed approach used the image read from NFC chips in the Chinese ID cards to compare against live faces. A base model is trained on an unconstrained face dataset (selfies), and then the knowledge is transferred to the target domain (ID card images). While the approach is not directly related to detecting presentation attacks on ID cards, the authors overcome it by accessing facial images stored in a chip within the ID document via NFC. However, such a solution cannot be employed for electronic ID cards enabled with storage.

3 | DATABASE

Due to the sensitive nature of ID cards, not many datasets are available for public research. In order to study and investigate the problem, we construct a new in-house database of ID cards consisting of German ID cards and residence permits as shown in Figure 1. The in-house database consists of video recordings of bona fide ID cards, printed ID cards and replayed ID cards using digital display considering various types of attacks. As the database is obtained from an operational scenario, the videos of ID cards in our newly collected dataset with the consent by the users have varying lighting conditions, background, and scenarios. The attacks represent in-the-wild settings with different kinds of printers and display devices representing a challenging set. Since the data collected are from an operational scenario, the videos are recorded from various android, apple smartphones and tablets. Further, the data are captured using a custom-made ID verification software, which records a video for 9 s with 30 frames per second.

Figure 2 illustrates the process of bona fide data capture where the frontal part of the ID card is typically captured at a resolution of 1280 × 720 pixels for 9 seconds at 30 frames per second. Each presentation corresponds to an average of 270 frames per ID verification attempt.

3.1 | Survey of existing databases

Due to the sensitive nature of data on the ID cards, the databases for research and commercial purposes are hard to obtain. So far there are two known databases apart from our

in-house database. We list the databases through an exhaustive search in Table 1. The Chilean database from Ref. [7] and Public-IvS [32] where the ID photos are cropped from the existing ID cards are present.

3.2 | Bona fide ID dataset

Through manual annotation of the ID card dataset, our newly created dataset consists of ID cards of 433 unique subjects resulting in 116,910 frames (9 s at 30 frames per second). Further, by eliminating the frames with no ID card present through manual verification, our dataset totals to 97,477 frames.

3.3 | Attack ID dataset

Similar to bona fide ID card presentation, the attack presentation in the dataset consists of video recordings from printed attacks and digital replay attacks. While the ID cards are printed using a variety of printers, the display/replay attacks correspond to ID cards displayed on smartphones and tablets. With careful manual annotation on the collected dataset, we identify 67 unique print attacks resulting in 18,090 frames (9 s at 30 frames per second). We further eliminate frames with no ID card in the frame resulting in 14,289 frames. The replay attack has 368 unique cases with 99,360 (9 s at 30 frames per second), and further curation of the dataset to eliminate frames with partial ID cards resulted in 80,388 frames.

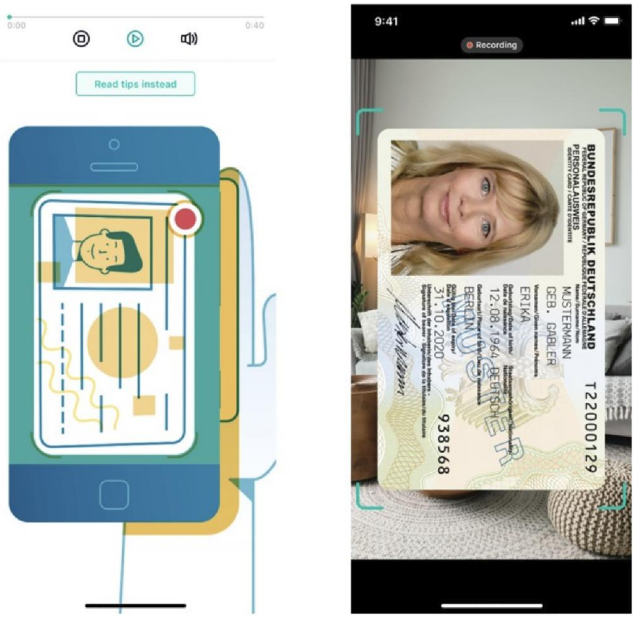


FIGURE 2 Illustration of data capture for ID cards using a custom-made ID verification software on mobile phone

4 | OUR APPROACH

Our proposed approach is based on observing minute cues differing in bona fide versus attack ID cards. Such cues can correspond to blur and moiré patterns introduced due to capture and printer related artefacts for attacks, while they tend to appear less frequently for bona fide presentations. We use this fact to our advantage by using pixel-wise supervision [20, 33] in addition to labelling the whole image with one single label as shown in Figure 3. We assert that such an approach can result in a better PAD approach by leveraging both pixel-level supervision and label-level supervision. Thus, our proposed approach repurposes pixel-wise supervision and label supervision for detecting presentation attacks on ID cards. The ID cards are pre-processed before providing it to the proposed network.

4.1 | Pre-processing using card alignment network

The ID cards are cropped from the frames using a card alignment without any processing. By ensuring no processing is done except cropping and resizing, no artefacts/artificial pixels are introduced.

4.2 | Proposed network

The proposed method follows a frame-wise approach in detecting the image as bona fide or attack. We train a densely connected network using both binary and pixel-wise

TABLE 1 Survey of the existing databases for Presentation Attack Detection (PAD) in ID cards

	Bona fide		Replay		Print			
Database	Videos	Frames	Videos	Frames	Videos	Frames	Publicly available	Remarks
Chilean ID Cards [7]	-	6588	-	24,778	-	6972	No	Bona fide images collected In varying lighting conditions
Public-IvS [32]	-	54,853	-	-	-	-	Yes	Only ID Photo of the ID cards available
Ours	433	10,677	366	79,926	67	14,279	No	Data collected in unrestricted environments *Can be used for evaluation of models by request.*

supervision. The model outputs a feature map and a predicted value. The output feature map can be considered to be the scores generated from the patches based on the filters in the network. Individual pixel patch is labelled as bona fide/attack as shown in Figure 3.

4.2.1 | DenseNet architecture

The motivation behind using DenseNet [22] is that in Convolutional Neural Network (CNN) as the model gets deeper the path from the input to the output layers grows immensely and the information to be conveyed is vanished. In DenseNet, the authors connect every layer with each other. This helps in the network in needing a smaller amount of parameters in comparison with standard CNNs. This results in filtering out redundant feature maps. The advantage of this network is that it does not take the sum of the output feature maps with input maps but *concatenates* them. This results in creation of DenseBlocks in which the dimensions of the feature maps are kept constant and the number of filters are varied. DenseBlocks with the combination of batch normalisation and downsampling are called the *Transition Layers*.

Following the motivation in Refs. [20, 33], our approach employs a simplified architecture with respect to the DenseNet architecture, which is modified with only two dense blocks and two transition blocks with a fully connected layer with sigmoid activation to produce the binary output. Such a modification also helps to prevent the over-fitting of the network under limited data availability. With the proposed architecture, the final layer has all the information from all the previous layers, including the input layer. In addition, a convolutional layer with a kernel size of 1×1 is introduced before the fully connected layer to generate the feature map for pixel-wise supervision. The feature map of size 14×14 is generated from this convolutional layer and is used to supervise the training of the network in a pixel-wise manner.

The architecture as shown in Figure 4 presents our backbone network with two transition blocks with a fully connected layer with sigmoid activation to produce the binary output. Further, we employ Binary Cross Entropy (BCE) for both pixel-wise and binary supervision, which can be represented as

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{BCE}^{pw} + (1 - \lambda) \cdot \mathcal{L}_{BCE}^{l} \quad (1)$$

where $\mathcal{L}_{BCE}^{pw}$ presents the loss computed based on pixel-wise supervision, $\mathcal{L}_{BCE}^{l}$ presents the label level loss computed based on binary output and λ is the regularisation parameter set to 0.5 in the experiments as motivated in Ref. [20].

The saved model has a size of 6 MB; the inference time for a video of 9 s with 30 fps with a skip rate of 3 on a Graphics Processing Unit (GPU) takes approximately 3 s to process a video.

FIGURE 3 Representation of the pixel-wise binary representation of the labels. Every pixel is given a binary label depending on the input data

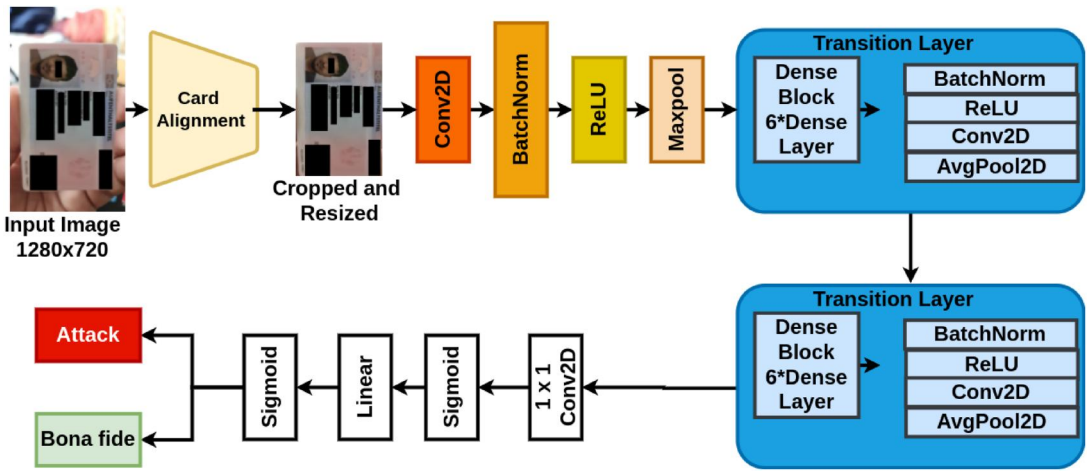


FIGURE 4 Proposed approach for ID card PAD using backbone of DenseNet architecture where each layer is connected to every other layer. For each layer the feature maps of all its previous layers are used as inputs to detect the presentation attacks on ID cards

4.3 | Hyper-parameters

We employ Adam optimiser [34] with a learning rate of 10^{-4} and a weight decay of 10^{-5} for training. Batch size is set to 32, and early stopping techniques are used to avoid over-fitting. The final score for each test image is computed by a binary output. Our proposed approach is trained on a GPU for 10 epochs. The performance is reported on the disjoint set unseen during training.

5 | EXPERIMENTS

We present the details of chosen baselines and metrics employed for measuring the PAD performance in this section along with the obtained results.

5.1 | Baseline methods—handcrafted

Due to the absence of direct literature on PAD for ID cards, we chose widely used face PAD approaches for our baseline results. Specifically, we utilised Binarised Statistical Image Features (BSIF) [35] and Co-occurrence of Adjacent Local Binary Patterns (CoALBP) with two different classifiers such as Support Vector Machines (SVM) and Collaborative Representation Classifier (CRC). While BSIF extracts features using statistically independent filters learnt on natural images, CoALBP is an extension of Local Binary Patterns (LBP), which utilises four immediate neighbours for calculating the co-occurrence frequency of two patterns resulting in a complex representation. Noting the better performance of joint features from CoALBP and Local Phase Quantization (LPQ), we resort to using CoALBP and LPQ on two different colour spaces of Hue-Saturation-Value and YCrCb [36]. Features from both chosen approaches are further represented using histograms for computational efficiency.

Specifically, in our experiments, we employ BSIF filters of size 5 × 5 filter with 12 bits, a combination of 9 × 9 with 12 bits, 11 × 11 filter and a combination of 11 × 11 with 12 bits with 17 × 17 with 12 bits filters are used to extract the texture descriptions of the bona fide and attacks data. Furthermore, SVM classifier is learnt using a Radial Basis Function (RBF) kernel to classify between the bona fide cases and the attacks.

5.2 | Baseline methods—deep learning

We further use three deep learning models based on ResNet [37], VGG16 [38] and MobileNet [39] using Imagenet pre-trained weights. All the three models learnt employ Adam optimiser [34] with sparse categorical cross-entropy as a loss function to train the models.

We trained the Deep models from scratch for 120 epochs. We observe that a random weight initialisation to the model does improve the model when compared to using the initialised weights from ImageNet as seen from Table 2.

5.3 | Performance evaluation metrics

We report performance of all methods using the standardized International Organization for Standardization/International Electrotechnical Commission 30,107-3:2017 metrics [40] defined for PAD using Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER). APCER is defined as the proportion of attack presentations species incorrectly classified as bona fide presentations in a specific scenario, and BPCER is defined as the proportion of bona fide presentations incorrectly classified as attack presentations in a specific scenario. We further report BPCER at APCER = 5% and APCER = 10% in line with studies on PAD along with the Detection-Equal Error Rate (D-EER%) as an indicative metric where APCER equals BPCER. In general, the lower the equal error rate value, the higher the accuracy of the biometric system. Finally, to complement our results reported at chosen thresholds of APCER, we present the Detection Error Trade-off (DET) curves for the convenience of the reader.

5.4 | Database splits

For all the results reported, we employ a disjoint training, validation and testing set following a disjoint split of 70% for training, 10% for validation and 20% for testing. The details of images in each split can be further obtained from Table 3.

5.5 | Results and discussion

As noted from the results presented in Table 2, we see that the SVM classifier learnt using handcrafted features perform better than the CRC classifier. The best performing approach of BSIF with a filter size of 5 × 5 with a bits 5 and SVM performs the best result with an EER of 6.78% while the BPCER at APCER = 5% equals 9.36%. However, the same approach deteriorates when learnt using CRC, and this can be easily explained by the linear classification as against the RBF kernel in SVM. Equally surprising are the results from deep models when the networks are learnt using pre-trained weights from ImageNet. The best performing deep model here is VGG16 whose EER equals 14.97%; however, the BPCER at APCER = 5% equals 29.22%.

In comparison, our proposed approach results in better EER equalling 3.24%. While the EER from the proposed approach is better than that of the other approaches, we note a superior BPCER at APCER = 5% and APCER = 10% equalling 2.64% and 2.09% indicating promising directions for using it in an operational scenario. Similar trends can also be noted in the DET curves presented in Figure 5.

We also study the proposed approach on data without frontalisation process. The obtained error rates as shown in Table 2 is slightly lower than the model with frontalisation. The key reason for this minor difference can be argued due to the

TABLE 2 Obtained results from proposed approach and baseline methods. The Equal Error Rate (EER) of our approach is lower than that of other approaches evaluated. *Note—The best performing handcrafted methods alone are reported based on best empirical trials

Approach	EER%	BPCER@5%	BPCER@10%
Handcrafted features + SVM			
BSIF 5 × 5	6.78	9.36	4.14
BSIF 9 × 9 + BSIF 11 × 11	14.51	19.8	16.25
BSIF 11 × 11 + BSIF 17 × 17	18.95	20.21	17.61
CoALBP + LPQ (HSV + YCrCb)	8.32	12.93	6.64
Handcrafted features + CRC			
BSIF 5 × 5	21.56	31.91	24.94
BSIF 9 × 9 + BSIF 11 × 11	14.07	16.69	14.87
BSIF 11 × 11 + BSIF 17 × 17	12.22	22.10	13.05
Deep models			
ResNet50	30.27	69.94	67.42
VGG16	14.97	29.22	19.04
MobileNet	14.32	50.29	28.74
Deep models—without imagenet initialisation—120 epochs			
ResNet50	17.84	20.47	11.77
VGG16	9.46	11.43	5.85
MobileNet	16.03	34.11	21.09
Proposed—Pixel-Wise supervision	**3.24**	**2.64**	**2.09**
Proposed—Pixel-Wise supervision (no frontalisation)	**2.22**	**1.83**	**1.67**
Proposed—Pixel-Wise supervision—Print	**5.05**	**5.15**	**2.72**
Proposed—Pixel-Wise supervision—Replay	**2.83**	**2.47**	**1.95**

TABLE 3 Details of the dataset collected in this work with a total of 866 unique ID cards from 866 unique subjects. The attacks are manually verified by ID verification experts as replay and print attacks

	Bona fide		Replay		Print	
Dataset	Videos	Frames	Videos	Frames	Videos	Frames
Train	303	74,148	256	56,513	47	9831
Validate	43	11,649	37	8435	6	1311
Test	87	20,870	73	14,978	14	3137

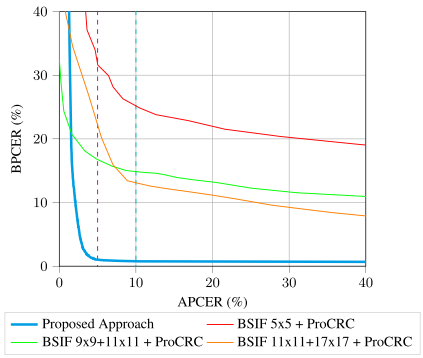


FIGURE 5 DET curves for ProCRC evaluated along with our proposed approach

presence of background. The presence of background increases the PAD performance by detecting attacks to a better extent.

5.6 | Individual error analysis

We present the error analysis for attacks individually for print and replay attacks as presented in Figure 6. We can observe that the EER of the replay attack is lower compared to that of the Print attack. The print attacks are more susceptible in the model to be falsely classified and contribute to the increased APCER. This can be further investigated by collecting more data samples and balancing out the attack dataset.

6 | ANALYSIS FOR EXPLAINABILITY OF PROPOSED PIXEL-WISE PAD

We further understand the proposed approach of pixel-wise binary PAD by utilising class activation maps (CAM). Through the CAM analysis, we assess the regions of significance with respect to every position of ID cards using the linear combinations of the activations obtained by last convolution layer and the output weights for the given class. The obtained CAM are further overlayed on the ID cards to

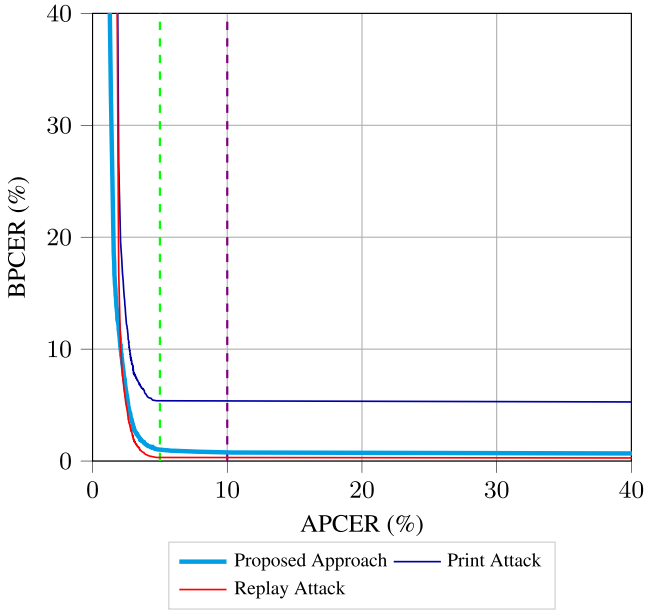


FIGURE 6 DET curves evaluated for our proposed approach with comparison to Print and Replay Attacks

illustrate the areas of significance in making the decisions as bona fide or attack presentations.

To further study the proposed approach in a detailed manner, we employ three different CAM such as GradCAM [41], ScoreCAM [42] and AblationCAM [43]. GradCAM employs the gradient loss with respect to the last convolution layer and computes the weighted combination of activation maps and ReLU for its respective feature maps. ScoreCAM [42] in contrast uses the weights of the activation maps from each forward pass and uses linear combination of the weights and activation maps. Using such an approach the ScoreCAM minimises the unstable nature of GradCAMs under the presence of noise. Similar to the GradCAM, the authors perform a linear combination of the target class to the target class score and each activation map. By applying ReLU to the resulting activation maps in ScoreCAM, the relevant areas used for learning the model can be obtained. AblationCAM [43] further finds weights based on the changes in the class scores by eliminating the specific regions of the feature maps corresponding to a class. In comparison to the GradCAM, AblationCAM is reported not to be sensitive to noisy images, saturation and vanishing gradients.

In GradCAM Figure 7a, we notice that the bona fide maps have uniform attention over entire image whereas the attention varies in the attack presentations. The varying attention intensity for attack presentations can be attributed to moiré patterns from videos and the artefacts left behind by the printers are highlighted in warmer and colder colour scales.

Further analysis indicates the need to focus on different regions either through a patch-wise approach or weighted patch-wise approach to be investigated in future works. While the outputs from GradCAM and ScoreCAM Figure 7b correlate to our intuition of using pixel-wise supervision for PAD, we note not so consistent results in AblationCAM Figure 7c.

7 | ABLATION STUDIES

We conduct ablation studies on the proposed network to demonstrate the effectiveness of the chosen parameters and the proposed framework. The results from ablation studies are presented in Table 4 and correspondingly we present the CAM analysis in Figure 8.

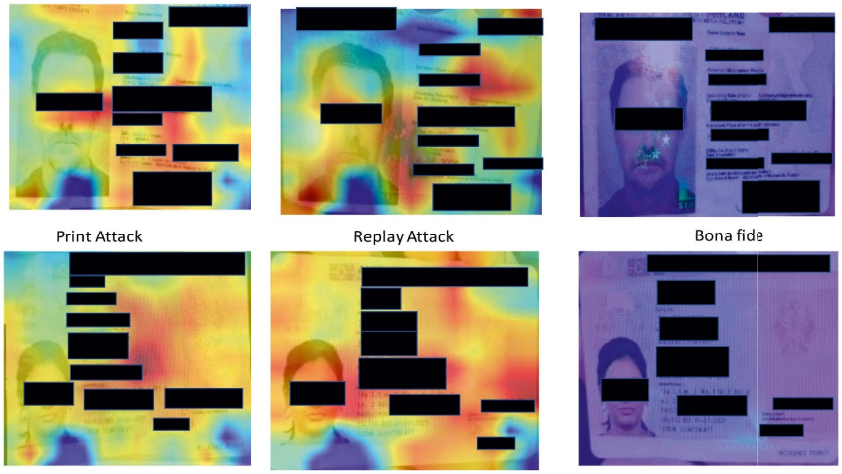
7.1 | Role of BCE loss

In order to study the effectiveness of employed Binary Cross-Entropy (BCE) loss, we replace the BCE to Mean Squared Error Loss (MSE) [44]. Cross-entropy calculates the score, which summarises the average difference between the true and the predicted probabilities for a given class, while the MSE is used as a regression metric where in the model is punished heavily if the variation of the predicted value is a large value. We note that such a loss does not outperform the BCE loss in our proposed approach as noted in Table 4. Despite not obtaining the lowest EER as against the best chosen configuration, we can anticipate an increase in performance if the loss is used in conjunction with the BCE and this aspect will be studied in future works.

7.2 | Contribution of independent losses

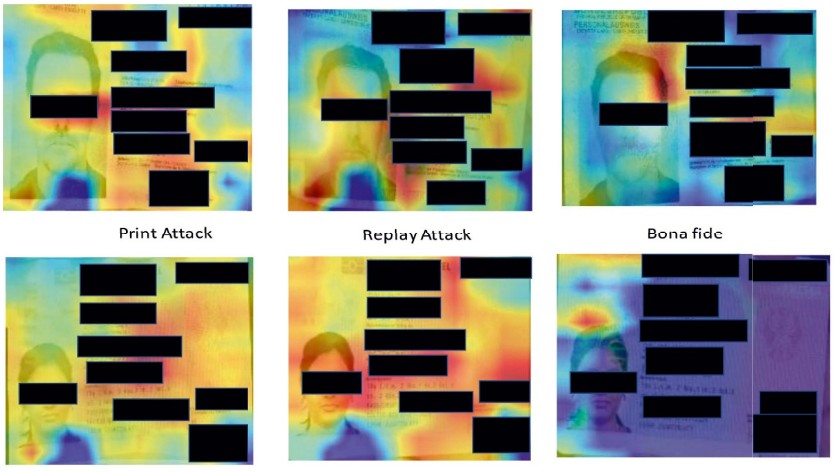
As noted in Equation (1), our proposed approach utilises pixel-wise supervision loss $\mathcal{L}_{BCE}^{pw}$ and binary label supervision loss $\mathcal{L}_{BCE}^{l}$. We thus study the contributions of each of the independent losses by varying the λ values in our loss function in Equation (1). By setting the $\lambda = 1$, we only make use of pixel-wise supervision loss $\mathcal{L}_{BCE}^{pw}$ and by setting $\lambda = 0$, we only utilise binary label supervision loss $\mathcal{L}_{BCE}^{l}$. Further, we also set different weights to λ to study the impact of various combinations. As noted from the Table 4, pixel-wise supervision alone results in a reasonable EER but with high BPCER at APCER = 5%, while the binary label supervision loss alone results in better EER and better BPCER at APCER = 5% and APCER = 10%. However, a balanced combination of both losses obtains the best EER and BPCER at both APCER = 5% and APCER = 10%.

The obtained results can further be correlated to the observations made through CAM analysis as shown in Figure 8 for the same set of images shown in Figure 7. As observed from GradCAM and ScoreCAM analysis, one can note that the network does not provide a clear decision on bona fide and attacks with just one of the losses. Such an observation further asserts our intuition of chosen network design.



(a) Gradient-weighted Class Activation Mapping (GradCAM)

(b) Score-Weighted Visual Explanations for Convolutional Neural Networks (ScoreCAM)



(c) Ablation-based Class Activation Mapping (AblationCAM)

FIGURE 7 CAM Analysis on bona fide and attack images

7.3 | Role of transition layer

Considering the superior performance of the proposed approach and to reduce the number of parameters of the model, we further conduct another analysis to study the feasibility of making network compact. We therefore analyse the proposed approach by eliminating one of the transition blocks—which contains six dense layers and the block of activation layers and average pooling blocks. As noted from the results presented in Table 4, removing one of the transition

TABLE 4 Analysis of results for various ablation studies that include different combination (weights) of losses, alternative loss and presence of single transition block in proposed approach

Approach	EER%	BPCER @ APCER = 5%	BPCER @ APCER = 10%
Combination of losses			
$\lambda = 0$	3.86	4.26	2.11
$\lambda = 0.2$	3.64	4.12	1.89
$\lambda = 0.9$	2.93	3.89	2.24
$\lambda = 1$	4.44	10.39	16.95
Proposed ($\lambda = 0.5$)	**3.24**	**2.64**	**2.09**
MSE Loss [44]	4.94	10.26	15.68
Single transition block	8.43	27.72	6.26

blocks deteriorates the performance of the proposed model by increasing both EER and BPCER. While removal of the block significantly decreases the training time and the GPU consumption, the performance of the model drops asserting our design choice. However, alternative strategies will be studied in future works to make the model compact while retaining the performance.

8 | FAILURE ANALYSIS OF PROPOSED APPROACH

In addition to other analysis noting a constant increase in BPCER at APCER<5% and increasing APCER at BPCER>1% as seen in Figure 5, we conduct another analysis to closely understand where the network fails. Specifically, we present a set of presentations where we obtain false negative (bona fides classified as attacks) and false positive (attacks classified as bona fides) in Figure 9. Visual analysis of the frames indicates that when the cropped frames are skewed and not aligned, the model fails to classify. To support our visual analysis, we provide CAM analysis of the misclassified frames, which show that the attention of the model is not focused to provide a clear decision.

First, we observe that the proposed approach is challenged when the presented ID card is not well aligned with the presence of fingertips in the background. We note that in such a case, the model rejects the bona fide attempt as attack attempt as shown in Figure 9a. We can further observe that the bright reflection on the face region in replay attacks as shown in Figure 9c leads the proposed approach to conclude the attack attempt as a bona fide attempt. Such errors lead to classification errors (APCER and BPCER) in the lower regions of the DET curve as shown in Figure 10. We further hypothesise that such errors can be easily mitigated by improved pre-processing approaches that can align the cards in a better manner and eliminate the artefacts arising out of illumination. Both these aspects will be studied in the follow-up works of this article.

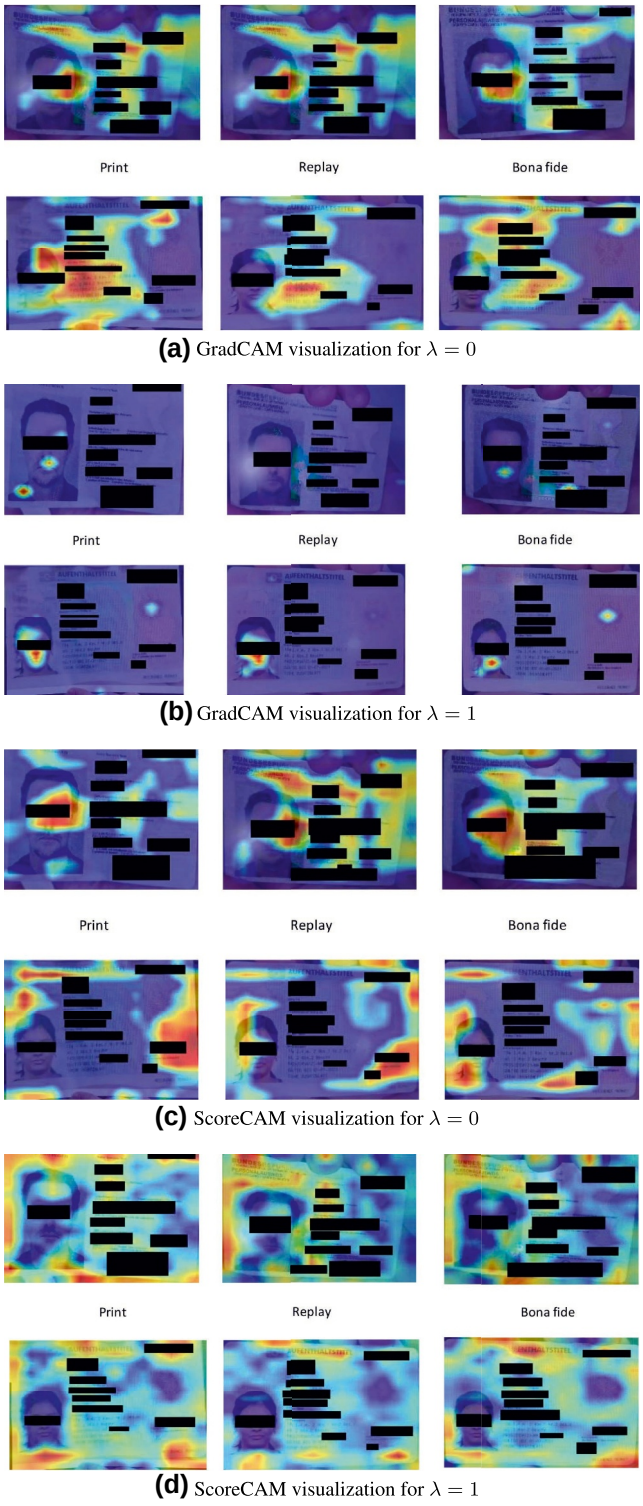


FIGURE 8 CAM Analysis on bona fide and attack images for the ablation studies

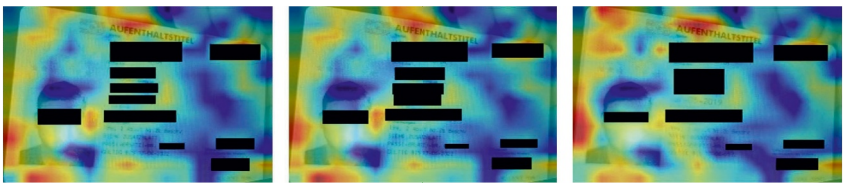
9 | LIMITATIONS OF OUR WORK

While the proposed approach provides a promising result, we note certain limitations in the current work. Specifically, in this work, the proposed approach is tested on a single in-house dataset limiting the robustness testing against different types

(a) Bona fide attempts classified as Attacks

(b) Class activation maps for Bona fide attempts classified as Attacks

(c) Replay attacks classified as Bona fide



(d) Class activation maps for replay attacks classified as Bona fide

FIGURE 9 Illustrative examples on which proposed approach failed and their corresponding class activation maps using GradCAM

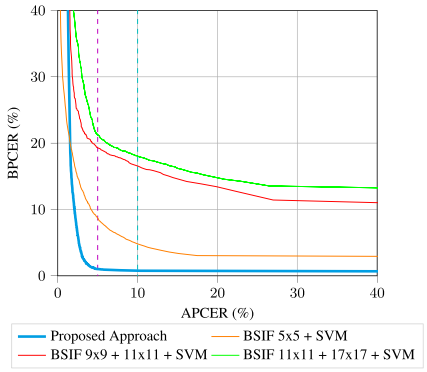


FIGURE 10 DET curves for SVM approaches evaluated along with our proposed approach

of ID cards. A diverse dataset of ID cards from different countries needs to be further investigated to identify the potential drawbacks of the proposed approach. Similarly, we have used pixel-wise supervision with DenseNet, and the generalisation of pixel-wise supervision against other networks are studied here; however, DenseNet has proven to generalise well over different architecture for the task of face PAD [45]. Further, the heavily imbalanced nature of the dataset with more bona fide can be seen as a case of anomaly detection, and such an approach has not been considered in our current work. Finally, the approach can be complemented by looking at the face region and other cues on the ID cards to make PAD further robust. All the limitations mentioned above shall be investigated in future works.

10 | CONCLUSION AND FUTURE DIRECTIONS

The verification of ID cards in unsupervised settings can be challenged by presentation attacks where an attack can simply employ digital copies or photocopied ID cards. To detect such presentation attacks on ID card verification systems, in this work, we have proposed a pixel-wise supervised learning paradigm using DenseNet to detect both printed and digital replay attacks. With a newly constructed in-house database obtained from an operational system consisting of 886 users with 433 bona fide, 67 print, 366 display attacks, future works in this direction can study the generalisation across ID cards and different attacks. Our model achieves an EER of 2.22% and lowest compared to the handcrafted features, pattern classifier and the deep learning models. The future works include in generalising the model to detect PAD on various kind of ID cards and ICAO documents, which can differ

significantly in appearance and design. As we focus on model generalisability and detection of edited photographs, personal information can also be checked. However, the NFC details on the cards cannot be read on all devices. Depending on the make and model of smartphone and issued year of the ID card, the attack detection can make use of data from NFC chips. We shall consider these aspects in the future works.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Raghavendra Mudgalgundurao https://orcid.org/0000-0002-1152-1018
Raghavendra Ramachandra https://orcid.org/0000-0003-0484-3956
Naser Damer https://orcid.org/0000-0001-7910-7895

REFERENCES

1. Mathias, K.: Forbes Digital Identity (2022). https://www.forbes.com/sites/forbestechcouncil/2022/01/07/the-next-evolution-of-digital-identity-in-2022/. [Online; accessed 2022-01-23]
2. Chris, B.: Biometrics Adoption (2021). https://www.biometricupdate.com/202201/stage-set-for-next-round-of-secure-access-control-biometrics-adoption. [Online; accessed 2022-01-23]
3. Shi, Y., Jain, A.K.D.: Matching id document photos to selfies. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems BTAS, pp. 1–8. IEEE (2018)
4. Shi, Y., Jain, A.K.: DocFace+: ID document to selfie matching. IEEE Transactions on Biometrics, Behavior, and Identity Science. 1(1), 56–67 (2019). https://doi.org/10.1109/tbiom.2019.2897807
5. Government, A.: Austrac (2022). https://www.austrac.gov.au/business/how-comply-and-report-guidance-and-resources/customer-identification-and-verification/kyc-requirements-covid-19. [Online; accessed 2022-01-23]
6. Bourlai, T., Ross, A., Jain, A.K.: Restoring degraded face images: a case study in matching faxed, printed, and scanned photos. IEEE Trans. Inf. Forensics Secur. 6(2), 371–384 (2011). https://doi.org/10.1109/tifs.2011.2109951
7. Gonzalez, S., Valenzuela, A., Tapia, J.: Hybrid two-stage architecture for tampering detection of chipless id cards. IEEE Transactions on Biometrics, Behavior, and Identity Science. 3(1), 89–100 (2020). https://doi.org/10.1109/tbiom.2020.3024263
8. Viedma, I., et al.: Fraud attack detection in remote verification systems for non-enrolled users. In: AI and Deep Learning in Biometric Security, pp. 239–256. CRC Press (2021)
9. Raghavendra, R., Raja, K.B., Busch, C.: Presentation attack detection for face recognition using light field camera. IEEE Trans. Image Process. 24(3), 1060–1075 (2015). https://doi.org/10.1109/tip.2015.2395951
10. Fang, M., et al.: Real masks and spoof faces: on the masked face presentation attack detection. Pattern Recogn. 123, 108398 (2022). https://doi.org/10.1016/j.patcog.2021.108398
11. Galbally, J., Marcel, S., Fierrez, J.: Biometric antispoofing methods: a survey in face recognition. IEEE Access. 2, 1530–1552 (2014). https://doi.org/10.1109/access.2014.2381273
12. Ramachandra, R., Busch, C.: Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Computing Surveys CSUR. 50(1), 1–37 (2017). https://doi.org/10.1145/3038924
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR'05, vol. 1, pp. 886–893. IEEE (2005)
14. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-Proceedings of the International Conference of Biometrics Special Interest Group BIOSIG, pp. 1–7. IEEE (2012)
15. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Process. 19(6), 1657–1663 (2010). https://doi.org/10.1109/tip.2010.2044957
16. Damer, N., Dimitrov, K.: Practical view on face presentation attack detection. In: BMVC. BMVA Press (2016)
17. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European Conference on Computer Vision, pp. 428–441. Springer (2006)
18. Zhou, P., et al.: Learning rich features for image manipulation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1053–1061 (2018)
19. Lee, W.H., Chou, C.M., Wang, S.W.: An nfc anti-counterfeiting framework for id verification and image protection. Mobile Network. Appl. 21(4), 646–655 (2016). https://doi.org/10.1007/s11036-016-0721-9
20. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. In: 2019 International Conference on Biometrics ICB, pp. 1–8. IEEE (2019)
21. Damer, N., et al.: Pw-mad: pixel-wise supervision for generalized face morphing attack detection. In: International Symposium on Visual Computing, pp. 291–304. Springer (2021)
22. Huang, G., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
23. Council, E.: Prado Eu Document Colections (2021). https://www.consilium.europa.eu/prado/en/prado-documents/DEU/index.html. [Online; accessed 2022-01-22]
24. Venkatesh, S., et al.: Face Morphing Attack Generation & Detection: A Comprehensive Survey. IEEE Transactions on Technology and Society (2021)
25. Czajka, A., Bowyer, K.W.: Presentation attack detection for iris recognition: an assessment of the state-of-the-art. ACM Computing Surveys CSUR. 51(4), 1–35 (2018). https://doi.org/10.1145/3232849
26. Sousedik, C., Busch, C.: Presentation attack detection methods for fingerprint recognition systems: a survey. IET Biom. 3(4), 219–233 (2014). https://doi.org/10.1049/iet-bmt.2013.0020
27. Sandler, M., et al.: Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
28. Nimble, N.: Datasets. (2016). https://www.nist.gov/itl/iad/mig/open-media-forensics-challenge
29. Dong, J., Wang, W., Tan, T.: Casia image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing, pp. 422–426. IEEE (2013)
30. Wen, B., et al.: Coverage—a novel database for copy-move forgery detection. In: 2016 IEEE International Conference on Image Processing ICIP, pp. 161–165. IEEE (2016)
31. Ng, T.T., Hsu, J., Chang, S.F.: Columbia Image Splicing Detection Evaluation Dataset. DVMM lab Columbia Univ CalPhotos Digit Libr (2009)
32. Zhu, X., et al.: Large-scale bisample learning on id versus spot face recognition. Int. J. Comput. Vis. 127(6), 684–700 (2019). https://doi.org/10.1007/s11263-019-01162-8
33. Fang, M., et al.: Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In: 2021 IEEE International Joint Conference on Biometrics IJCB, pp. 1–8. IEEE (2021)
34. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (2014). arXiv preprint arXiv:14126980

35. Kannala, J., Rahtu, E.B.: Binarized statistical image features. In: Proceedings of the 21st International Conference on Pattern Recognition ICPR2012, pp. 1363–1366. IEEE (2012)
36. Heikkila, J., Ojansivu, V.: Methods for local phase quantization in blur-insensitive image analysis. In: 2009 International Workshop on Local and Non-local Approximation in Image Processing, pp. 104–111. IEEE (2009)
37. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
38. Qassim, H., Verma, A., Feinzimer, D.: Compressed residual-vgg16 cnn model for big data places image recognition. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference CCWC, pp. 169–175. IEEE (2018)
39. Howard, A.G., et al.: Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017). arXiv preprint arXiv:170404861
40. Biometrics, I.J.S.: Iso/iec 30107-3. Information Technology - Biometric Presentation Attack Detection - Part 3: Testing and Reporting (2017)
41. Selvaraju, R.R., et al.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
42. Wang, H., et al.: Score-cam: score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 24–25 (2020)
43. Ramaswamy, H.G., et al.: Ablation-cam: visual explanations for deep convolutional network via gradient-free localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 983–991 (2020)
44. Sammut, C., Webb, G.I.: Mean squared error. In: Encyclopedia of Machine Learning, p. 653. Springer (2010)
45. Fang, M., et al.: Partial attack supervision and regional weighted inference for masked face presentation attack detection. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition FG 2021, pp. 1–8 (2021)

How to cite this article: Mudgalgundurao, R., et al.: Pixel-wise supervision for presentation attack detection on identity document cards. IET Biome. 11(5), 383–395 (2022). https://doi.org/10.1049/bme2.12088