IDIAP RESEARCH REPORT

EDGEDOC: HYBRID CNN-TRANSFORMER MODEL FOR ACCURATE FORGERY DETECTION AND LOCALIZATION IN ID DOCUMENTS

Anjith George[a] Sébastien Marcel

Idiap-RR-08-2025

AUGUST 2025

[a] Idiap Research Institute

EdgeDoc: Hybrid CNN-Transformer Model for Accurate Forgery Detection and Localization in ID Documents

Anjith George and Sébastien Marcel
Idiap Research Institute, Switzerland
{anjith.george,sebastien.marcel}@idiap.ch

Abstract

The widespread availability of tools for manipulating images and documents has made it increasingly easy to forge digital documents, posing a serious threat to Know Your Customer (KYC) processes and remote onboarding systems. Detecting such forgeries is essential to preserving the integrity and security of these services. In this work, we present EdgeDoc, a novel approach for the detection and localization of document forgeries. Our architecture combines a lightweight convolutional transformer with auxiliary noiseprint features extracted from the images, enhancing its ability to detect subtle manipulations. EdgeDoc achieved third place in the ICCV 2025 DeepID Challenge, demonstrating its competitiveness. Experimental results on the FantasyID dataset show that our method outperforms baseline approaches, highlighting its effectiveness in real-world scenarios. Project page : https://www.idiap.ch/paper/edgedoc/

1. Introduction

The widespread adoption of digital KYC processes in financial services has introduced new security risks, as forged identity documents can be injected or physically replayed to bypass verification systems. Advances in image generation [1] and editing have made such forgeries increasingly realistic and harder to detect, especially when involving subtle text manipulations. Although recent methods [11, 13, 14, 16–18] have improved manipulation detection, many overlook fine-grained document-level forgeries.

A major challenge in the domain of forgery detection is the lack of model generalization. Identity documents exhibit substantial variation in design across different regions, making it difficult to develop a single model capable of effectively generalizing across all layout types. Moreover, the nature of forgery attacks can differ significantly: some may involve alterations to textual content, others may target facial images, and some may modify both. Detecting forgeries becomes particularly challenging when the tampered region is relatively small. An additional limitation is the requirement for large-scale datasets to train robust models. However, the sensitive and personally identifiable nature of identity documents poses significant constraints on data collection and sharing, limiting the availability of comprehensive and realistic training datasets.

Motivated by the significance of the problem and its inherent challenges, we introduce EdgeDoc, a novel method for document manipulation detection. The proposed model employs a lightweight hybrid architecture that integrates convolutional and transformer-based components, enabling simultaneous classification and forgery localization. Designed to perform effectively with a limited number of training samples, EdgeDoc achieves competitive results on a public benchmark challenge.

2. Proposed Method

The proposed method, EdgeDoc, is a hybrid model that combines the strengths of the TruFor framework [4] with a custom lightweight architecture inspired by EdgeFace [3]. Given the limited availability of training data, training a model from scratch is impractical. To address this, we leverage the NoisePrint representation extracted via the TruFor pipeline, which serves as a source of localized anomaly cues. This NoisePrint is fused with the original image to enable patch-wise interaction within a convolutional-transformer architecture, facilitating both manipulation detection and localization. The details of the proposed approach are presented in the following subsections.

2.1. Device fingerprint extraction

In [2], the authors introduced NoisePrint, a neural network designed to extract a camera model fingerprint. The model is trained to suppress scene-related artifacts while enhancing camera model-specific patterns. Training is performed using a Siamese architecture, where image patch pairs from the same camera are treated as positive examples, and those from different cameras as negative examples.

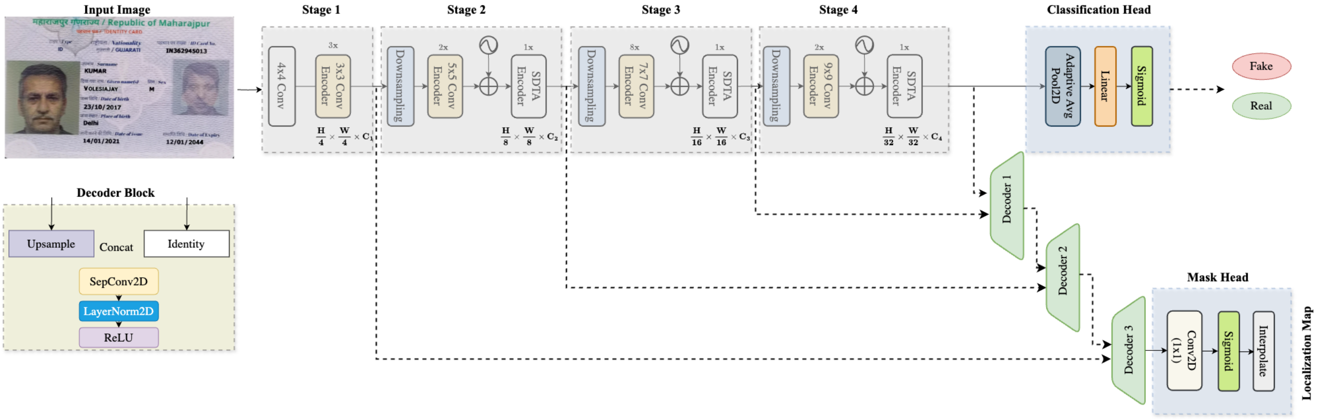

Figure 1. Architecture of the proposed EdgeDoc framework

Building on this, the TruFor framework [4] improved the method by incorporating a transformer-based fusion module that combines RGB information with an enhanced version of NoisePrint (called NoisePrint++). Forgery detection is approached as the identification of deviations from the expected regularity in the image. The framework outputs an integrity score, a localization map, and a confidence score, helping to identify potentially error-prone regions with higher precision.

2.2. Convolutional-Transformer Hybrid Network

EdgeFace [3] demonstrated the effectiveness of convolutional-transformer hybrid architectures for face recognition tasks, building on the EdgeNeXt framework [9]. These architectures combine the local inductive biases of convolutional layers with the global modeling capabilities of transformers, all within a compact and computationally efficient design. Motivated by their balance of performance and efficiency, we adopt a similar lightweight hybrid architecture in the development of our model.

2.3. EdgeDoc Architecture

Our proposed architecture, EdgeDoc, is based on the XXS variant of the EdgeNeXt backbone. It extracts multi-scale feature maps from various stages of the network, which are then fed into a custom decoder structured in a U-Net style. The architecture of EdgeDoc is shown in Fig. 1. The decoder is composed of upsampling blocks, each consisting of depthwise separable 2D convolutions, followed by 2D Layer Normalization and ReLU activations.

For classification, we utilize a bottleneck head comprising global average pooling and fully connected layers. The final segmentation mask is generated via a pointwise (1×1) convolution applied to the decoder output.

2.4. Training Details

The input to the network comprises two channels: the green channel of the ID image and the NoisePrint feature map. For the classification task, binary cross-entropy (BCE) loss was employed. Localization was optimized using a composite loss function combining BCE and Dice loss [10]. A weighting factor of $\lambda = 3.0$ was applied to the mask loss component during training. The total loss function ($\mathcal{L}_{\text{total}}$) is defined as follows:

$$\mathcal{L}_{\text{cls}} = \text{BCE}(y_{\text{cls}}, \hat{y}_{\text{cls}}) \tag{1}$$

$$\mathcal{L}_{\text{mask}} = \text{BCE}(y_{\text{mask}}, \hat{y}_{\text{mask}}) + \text{Dice}(y_{\text{mask}}, \hat{y}_{\text{mask}}) \tag{2}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\text{mask}} \tag{3}$$

where $\hat{y}_{\text{cls}}$ is the output from the classification head and $\hat{y}_{\text{mask}}$ the output from the mask head, y_{cls} and y_{mask} denote corresponding ground truths.

The model was trained using the AdamW [8] optimizer with a weight decay of 5×10^{-4} and a batch size of 1. The initial learning rate was set to 3×10^{-4} and decayed according to a cosine annealing schedule over 20 epochs. The model achieving the lowest validation loss during training was selected for final evaluation. All experiments were conducted on an NVIDIA RTX 3090 GPU.

3. Experiments

This section presents the experimental setup, baseline methods, and the corresponding results.

Dataset: FantasyID [5] is a high-quality dataset developed to support research in document forgery and presentation attack detection within biometric KYC systems. It consists of two categories: bonafide identity cards and attack samples. The bonafide subset includes 262 synthetically generated fantasy ID cards featuring randomized personal information and real facial images sourced from pub-

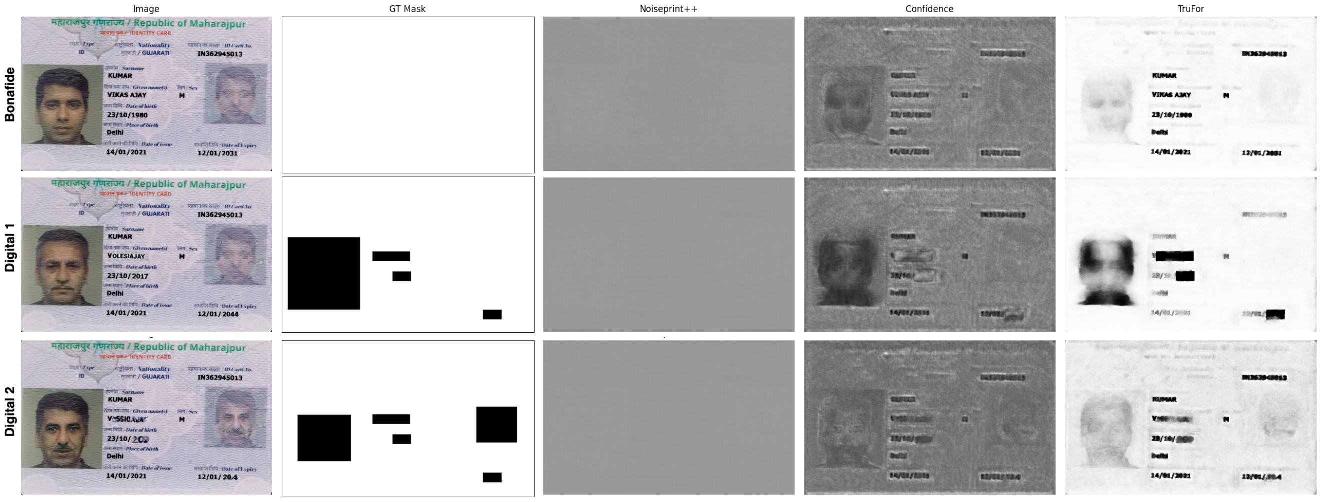

Figure 2. Sample and Ground Truth from Fantasy ID dataset, together with the NoisePrint++, Confidence and Localization results from TruFor

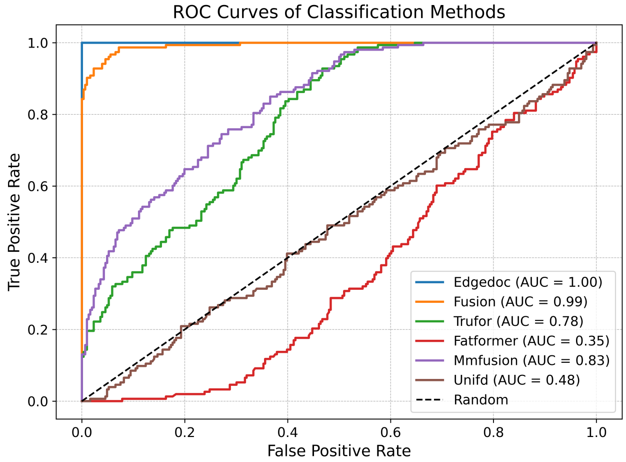

Figure 3. ROC curves for various methods from the literature on the public validation set of FantasyID dataset, along with our proposed EdgeDoc and Fusion approaches.

lic datasets. These cards were printed on plastic using a commercial ID card printer and captured under realistic conditions using three different imaging devices, resulting in 786 genuine images. The attack subset includes both digital manipulations created by altering facial and textual content using state-of-the-art generative models and printed manipulations, where digitally forged cards were reprinted and re-captured to simulate more sophisticated presentation attacks. Samples for bonafide and attacks are shown in Fig. 2. This dual-type attack design closely reflects real-world adversarial scenarios, offering a comprehensive and challenging benchmark for evaluating the robustness of document forgery detection algorithms. In this work, we use only the public training and validation set for the experiments.

Metrics: We report a comprehensive set of performance metrics for the binary classification task. The accuracy metric captures overall correctness but can be misleading under class imbalance. To address this, we include the weighted F1-score, which balances precision and recall by class weight. ROC AUC provides a threshold-independent view of true- vs. false-positive trade-offs, while the Matthews Correlation Coefficient (MCC) offers a balanced summary measure, robust even under severe skew. Together, these metrics offer a detailed evaluation of model performance across different conditions.

Baselines: We utilize four state-of-the-art algorithms for binary manipulation detection in images: TruFor [4], MMFusion [15], UniFD [12], and FatFormer [7]. We utilize the pretrained models released by the respective authors for our experiments.

Experimental Results: We trained the proposed EdgeDoc model using the training set of the Fantasy ID dataset and evaluated its performance on the corresponding validation set. In addition, we assessed several off-the-shelf baseline methods, including TruFor, for comparative analysis. The results of these evaluations are summarized in Table 1, where EdgeDoc demonstrates superior performance compared to all other methods. Receiver Operating Characteristic (ROC) curves for the baselines are presented in Figure 3. Furthermore, we explored a fusion of EdgeDoc and TruFor using a weighted combination, which also yielded competitive results.

ICCV 2025 DeepID Challenge Submission The ICCV 2025 DeepID Challenge [6] represents the first competition focused on detecting synthetic manipulations, specifically

Table 1. Performance on the public validation set of Fantasy ID

Model	**Accuracy**	**F1 (weighted)**	**Precision**	**Recall**	**ROC AUC**	**MCC**
Fatformer	0.34	0.21	0.33	0.93	0.35	-0.06
Mmfusion	0.69	0.59	1.00	0.08	0.83	0.23
Unifd	0.33	0.17	0.33	1.00	0.48	0.00
TruFor [4]	0.71	0.70	0.59	0.44	0.78	0.32
EdgeDoc	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Fusion(EdgeDoc, TruFor)	0.95	0.95	0.99	0.87	0.99	0.90

injection attacks as opposed to traditional presentation attacks in identity documents. As part of the challenge, the organizers released the train and development partitions of the Fantasy ID dataset along with corresponding ground-truth labels.

Given the limited availability of publicly accessible training data, we utilized a pretrained TruFor model to extract NoisePrint maps, which provide localized cues indicative of potential manipulations. These maps, together with the original images, were used as inputs to train our custom EdgeDoc model. EdgeDoc is designed to produce both a binary classification score and a segmentation mask, thereby enabling simultaneous detection and localization of forgeries. The model was trained solely on the public training subset of the Idiap Fantasy ID dataset.

For inference, we applied a fusion strategy combining the outputs of both EdgeDoc and TruFor, specifically their classification scores and localization masks to generate a final prediction score. The performance results, including those for the individual models and their fusion, as reported on the official competition leaderboard, are summarized in Table 2. While EdgeDoc and TruFor individually exhibit limited performance on the private test set, their fusion significantly improves generalization, demonstrating strong robustness to previously unseen manipulation scenarios.

Table 2. Performance Metrics from the Competition Leaderboard on the Fantasy ID and Private Test Datasets

Model	**F1 on Fantasy**	**F1 on Private**	**Aggregate F1**
EdgeDoc	0.43	0.66	0.59
TruFor [4]	0.81	0.66	0.71
Fusion (EdgeDoc, TruFor)	**0.96**	**0.71**	**0.79**

4. Conclusions

In this work, we present EdgeDoc, a lightweight framework for document forgery detection that leverages both original images and their corresponding NoisePrint representations. The proposed method demonstrated superior performance compared to other models on the development set of the Fantasy ID dataset. Our submission to the ICCV 2025 DeepID Challenge secured third place in the detection track, highlighting the effectiveness and competitiveness of the approach. We believe that with access to larger and more diverse training data, the performance of EdgeDoc can be further improved. To support future research and development, we will release the source code publicly.

5. Acknowledgments

This research was funded by the Swiss Center for Biometrics Research and Testing.

References

[1] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022. 1

[2] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019. 1

[3] Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Ketan Kotwal, and Sébastien Marcel. Edgeface: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):158–168, 2024. 1, 2

[4] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 1, 2, 3, 4

[5] Pavel Korshunov, Amir Mohammadi, Vidit Vidit, Christophe Ecabert, and Sébastien Marcel. Fantasyid: A dataset for detecting digital manipulations of id-documents. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2025. 2

[6] Pavel Korshunov, Vidit, Amir Mohammadi, Christophe Ecabert, Nevena Shamoska, Sébastien Marcel, Zeqin Yu, Ye Tian, Jiangqun Ni, Lazar Lazarevic, Renat Khizbullin, Anastasiia Evteeva, Alexey Tochin, Aleksei Grishin, Anjith George, Daniel DeAlcala, Tamás Endrei, Javier Mu noz Haro, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, Julian Fierrez, György Cserey, Hardik Sharma, Sachin Chaudhary, Akshay Dudhane, Praful Hambarde, Amit Shukla, Prateek Shaily, Jayant Kumar, Ajinkya Hase, Satish Maurya, Mridul Sharma, and Pallav Dwivedi. Deepid challenge of detecting synthetic manipulations in id documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025. 3

[7] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024. 3

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[9] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European conference on computer vision*, pages 3–20. Springer, 2022. 2

[10] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 2

[11] Javier Muñoz-Haro, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, and Julian Fierrez. Exploring a patch-wise approach for privacy-preserving fake id detection. *arXiv preprint arXiv:2504.07761*, 2025. 1

[12] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 3

[13] Alvaro Sanchez, Juan M Espín, and Juan E Tapia. Few-shot learning: Expanding id cards presentation attack detection to unknown id countries. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2024. 1

[14] Juan E Tapia, Naser Damer, Christoph Busch, Juan M Espin, Javier Barrachina, Alvaro S Rocamora, Kristof Ocvirk, Leon Alessio, Borut Batagelj, Sushrut Patwardhan, et al. First competition on presentation attack detection on id card. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024. 1

[15] Konstantinos Triaridis and Vasileios Mezaris. Exploring multi-modal fusion for image manipulation detection and localization. In *International conference on multimedia modeling*, pages 198–211. Springer, 2024. 3

[16] Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, and Ping Luo. Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4233–4245, 2025. 1

[17] Lixin Wang, Zhenjiang Li, and Wenqi Zhao. Research on identity document image tampering detection based on texture understanding and multistream networks. *Journal of Electronic Imaging*, 34(4):043018–043018, 2025.

[18] Lin Zhao, Changsheng Chen, and Jiwu Huang. Deep learning-based forgery attack on document images. *IEEE Transactions on Image Processing*, 30:7964–7979, 2021. 1