Ingénierie des Systèmes d'Information
Vol. 30, No. 8, August, 2025, pp. 2077-2084
Journal homepage: http://iieta.org/journals/isi

ID Card Spoofing Detection Using Frequency Features and CNNs

Tat Thang Nguyen*, Minh Manh Vo

Faculty of Information Technology, Posts and Telecommunications Institute of Technology - PTIT, Hanoi 100000, Vietnam

Corresponding Author Email: thangnt@ptit.edu.vn

https://doi.org/10.18280/isi.300813

Received: 27 June 2025
Revised: 19 August 2025
Accepted: 25 August 2025
Available online: 31 August 2025

Keywords:
identity card spoofing, convolutional neural network, deep convolutional network, Fourier Transform, cross-attention mechanism

ABSTRACT

This paper proposes a lightweight and robust approach for detecting spoofed ID card images by integrating convolutional neural networks (CNNs) with frequency-domain analysis. The model adopts a dual-branch design: one branch processes the original RGB image, while the other takes a frequency-enhanced version produced using a high-pass Fast Fourier Transform (FFT) filter. Both branches use the same architecture: the first seven convolutional layers of the VGG16 backbone but each branch has its own parameters. The two streams are merged by a multi-head cross-attention fusion module, which aligns and integrates the complementary cues from both branches more effectively, followed by a classification module for "genuine" vs. "spoof". The method is evaluated on the "or" and "re" subsets of the Document Liveness Challenge 2021 dataset (DLC-2021). On these subsets, the model attains precision of 93.64%, recall of 88.90% and accuracy of 91.68%, significantly outperforming baseline models. The implementation remains computationally efficient, requiring about 0.120 s per image on an Intel Xeon 2.20 GHz (x86-64) CPU. The approach achieves a favorable trade-off by combining high detection accuracy with a compact model size. These results highlight the benefit of exploiting both spatial and frequency features to enhance the reliability of electronic identity verification systems.

1. INTRODUCTION

The use of identity documents for electronic identity verification has expanded rapidly in recent years. Electronic Know Your Customer (eKYC) systems are now widely deployed in digital banking, e-wallets, e-commerce, and public online services. By 2023, approximately 11.9 million bank accounts in Vietnam had been opened through electronic identification, with most banks already adopting eKYC procedures [1, 2]. Compared to traditional face-to-face verification, eKYC offers a faster, more convenient, and cost-effective alternative.

Figure 1. Example of the electronic identity verification process with a genuine ID card (left) and a spoofed ID card (right)

However, this rapid adoption has also introduced new security risks. In 2023 alone, Vietnamese citizens lost an estimated VND 8,000-10,000 billion (USD 300-400 million) to cyber scams, with financial fraud accounting for 91% of these cases [3, 4]. Among these threats, identity spoofing has emerged as a critical challenge. A common technique involves indirect image presentation, where an attacker displays an ID card image on a digital device (e.g., smartphone or tablet) and presents it to the camera during live capture. As illustrated in Figure 1, such attacks can deceive eKYC systems into accepting spoofed inputs as genuine documents. These vulnerabilities not only lead to substantial financial losses but also weaken public trust in digital services.

To address this problem, advanced image analysis techniques are required. Recent progress in deep learning, particularly convolutional neural networks (CNNs), has shown strong potential in image classification, while Fourier-based methods have long been effective in revealing frequency-domain characteristics [5, 6]. Combining these approaches offer a promising path toward more accurate and reliable spoof detection.

In this study, we propose a dual-branch CNN model that integrates a VGG16 backbone with Fast Fourier Transform (FFT) processing [7]. The first branch captures spatial features from RGB images, while the second emphasizes frequency-enhanced characteristics that are often indicative of spoofing artifacts. The fused features are then classified into either "genuine" or "spoof." Unlike prior approaches that primarily focus on speed or lightweight design, our method prioritizes accuracy and parameter efficiency, demonstrating that higher detection performance can be achieved even when additional frequency-domain processing increases computational load.

The remainder of this paper is organized as follows: Section 2 reviews related work on document spoofing detection.

Section 3 presents the proposed model architecture. Section 4 describes the training and evaluation methodology. Section 5 reports and discusses experimental results, and Section 6 concludes with future directions.

2. RELATED WORK

Polevoy et al. [8] released the benchmark dataset called DLC-2021 for document liveness, with subsets “or” (original captures) and “re” (screen-recaptures), along with fixed train and test lists and standard metrics for fair comparison. In addition, the authors provided a ResNet-50 baseline [9, 10] to distinguish between genuine and spoofed images. The model used pre-trained weights from the ImageNet dataset [11], with 49 out of 50 layers frozen. The final softmax layer was replaced with a binary output layer. The binary cross-entropy loss function was applied, and the Adam optimizer was used with a learning rate of 0.1. For both training and evaluation, the baseline operated on a fixed 224 × 224 crop centered at the annotated card centroid. The model was trained and evaluated on the DLC-2021 and MIDV-2020 [12] datasets. Specifically, the training set included 19,543 spoofed images and 25,980 genuine images, while the evaluation set comprised 11,009 spoofed and 16,264 genuine images. The results showed strong performance, with a precision of 85.89%, recall of 89.03%, and accuracy of 89.67%.

Kunina et al. [13] introduced a boundary-focused screen-recapture detector that enhances narrow strips along the document edge and applies a fast Hough transform to expose stripe-like moiré leaking across the border - treated as a replay artifact rather than background texture. On DLC-2021, the reported precision is 93.5%, recall is 90.2%, and accuracy is 92.1% baseline test split, while running 1.5-2 times faster than the DLC-2021 baseline.

Markham et al. [14] used four generative models to synthesize spoofed identity card images, which were then used to train a MobileNetV2 network [15] for genuine-vs-spoof classification. The test dataset was created by aggregating public datasets such as MIDV-2020 [12] and DLC-2021 [8]. On the combined test set including both genuine and screen-recapture spoofed images, the model achieved an Equal Error Rate (EER) of 5.8%. Precision, recall, and accuracy were not reported.

Al-Ghadi et al. [16] introduced IDTrust - a deep learning framework for ID card liveness that separates originals from scanned and printed copies without requiring a reference template. Two proposed variants are DeepQD, which uses a CNN encoder plus a binary head, and GuidedDeepQD, which prepends an FFT-based band-pass filtering step to emphasize background and guilloche patterns degraded by scanning before feeding the CNN. The proposed model reached near-ceiling accuracy on MIDV-2020 [12].

Research on synthetic datasets for identity and travel documents has been conducted by Boned et al. [17] and Guan et al. [18], who created large-scale, privacy-preserving resources that capture diverse document types to support training and evaluation of fraud detection systems.

Zhang et al. [19] provided a systematic review of AI-based methods for identity fraud detection, summarizing 43 studies, categorizing existing approaches, and outlining their advantages, limitations, and open challenges.

In a broader context of face spoofing detection, Zeng et al. [20] proposed a dual-branch CNN model based on EfficientFormerV2 [21, 22]. One branch processed the original face image, while the other received a version preprocessed with a high-pass filter (Gaussian) and the Fourier Transform. A multi-head cross-attention module was then used to fuse features from both branches. On the Face Anti-Spoofing Challenge@CVPR2023 dataset, this model achieved an Average Classification Error Rate (ACER) of 6.22% and ranked fourth in the competition. Yu et al. [23] approached face spoofing detection by replacing conventional convolutional layers with Central Difference Convolution, combining both intensity and gradient information at each pixel. This architecture performed well, achieving a low ACER of 0.2% on the OULU-NPU dataset under Protocol-1 [24] and an HTER of 6.5% in cross-dataset evaluation on CASIA-MFSD [25]. Shinde and Raundale [26] propose a two-stage face-spoofing detection framework: (i) CNN-based presentation attack classifier, which discriminates between bona fide faces and various spoofing media (printed photographs, posters, ...) and (ii) liveness verification, combining an eye‐blink detection module with a lip‐motion analysis module to ensure that the presented face originates from a live subject rather than a static image or screen.

These spoofing detection methods tend to focus on using cutting-edge deep learning architectures, which are often large in scale and complex in structure. However, they have not placed much emphasis on exploring simpler, more stable CNN architectures that are easier to deploy and fine-tune - particularly for systems with real-time processing constraints and limited hardware resources. Therefore, in this study, we focus on utilizing the basic VGG16 CNN model, and enhancing its effectiveness by introducing a novel architecture incorporating the Fourier Transform. We also design an efficient training strategy and apply additional image processing techniques to improve the accuracy of identity card spoofing detection.

3. PROPOSED SYSTEM ARCHITECTURE

3.1 Overview

In the frequency-domain analysis of images, the spectrum is typically divided into low-frequency and high-frequency components, each carrying different information. The low-frequency component represents areas with slowly varying, relatively uniform intensity such as the walls of a room or a cloudless sky in an outdoor scene. These values concentrate around the center of the image’s frequency spectrum. In contrast, the high-frequency component appears near the edges of the spectrum due to abrupt intensity changes. High frequencies usually correspond to sharp edges, textural details, or noise [5, 27].

Spoofed identity document images often exhibit distinct characteristics compared to genuine ones. First, such images frequently contain hard edges, such as the borders of a phone screen or the frame of a computer monitor. These features can be effectively captured and classified using the VGG16 model. Second, spoofed images often include visual artifacts like horizontal stripes or noise patterns originating from LCD or LED screens. These subtle details can be amplified using a high-pass frequency filter via the Fourier Transform (as illustrated in Figure 2). Third, indirect image capture through an intermediary screen typically results in blurrier images

compared to direct captures. This blur is another key feature that can be distinguished using high-frequency filtering.

Based on these observations, we propose a dual-branch architecture named VGG16 + FFT. The first branch processes the original image (referred to as the main branch), while the second branch processes the image after applying the Fourier Transform (referred to as the FFT branch). The outputs from both branches are fused using a multi-head cross-attention module to effectively integrate information. Finally, a classification module produces the binary output (Figure 3).

Figure 2. Screen stripe patterns enhanced using a high-pass frequency filter with a mask radius of 8

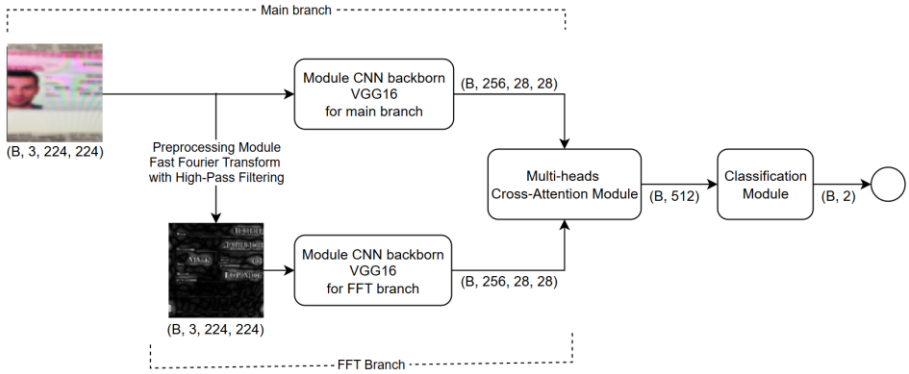

Figure 3. Proposed architecture of the VGG16 + FFT model

The network input is a batch of B images (in RGB format with 3 color channels), where pixel values are normalized from the range [0,255] to [0,1]. For the main branch, the input is further normalized using ImageNet standards, resized to shape (B,3,224,224) and converted into tensor format. This tensor is then passed through the VGG16 backbone, retaining only the first four convolutional blocks which is equivalent to the first 7 layers of the model to reduce overall network complexity. The output of this branch is a feature tensor of shape (B,256,28,28). In the FFT branch, the input undergoes a preprocessing module that applies the Fast Fourier Transform (FFT) combined with a high-pass filtering mechanism. This highlights high-frequency components in the image, resulting in a tensor of shape (B,3,224,224). This process enhances features useful for distinguishing genuine images from spoofed ones. The processed images are then passed through a VGG16-based backbone identical to that used in the main branch. The output of this branch is also a feature tensor of shape (B,256,28,28). Once the two feature tensors are obtained, they are fed into a multi-head cross-attention module, which enables the model to learn interactions and correlations between spatial-domain features (from the main branch) and high-frequency enhanced features (from the FFT branch). The output is a fused tensor of shape (B,512) integrating information from both branches and ready to be passed into a fully connected classification head. The binary classification module takes the 512-dimensional fused feature tensor as input. It is processed through three fully connected layers (optionally including batch normalization), followed by a softmax activation function to produce a probability distribution over the two output classes: "genuine" and "spoof".

3.2 Preprocessing module: FFT with high-pass filtering

To support the FFT branch, we design a preprocessing module consisting of the following steps:

Step 1: Cropping and normalizing the region of interest

Instead of processing the full frame, the pipeline first extracts two square crops centered at the annotated card centroid: a large crop $L_1 = 1120 \times 1120$ and a smaller crop $L_2 = 448 \times 448$. Both crops are resized to 224×224. The image cropped with L_1 is routed to the spatial branch (VGG16), while the image cropped with L_2 is fed to the frequency branch (FFT). The choice of L_1 and L_2 follows the protocol: L_1 provides broader contextual content for robust spatial cues, whereas L_2 concentrates on micro structures (e.g., moiré), thereby, helping the model learn more distinctive features to differentiate between genuine and spoofed ID images.

Step 2: FFT, frequency shift and filtering

L_2 cropped image is converted to grayscale and formatted as a tensor $x_{gray} \in \mathbb{R}^{B\times1\times224\times224}$. Apply FFT and center DC with $FFTShift$.

$$X = FFT(x_{gray}) \tag{1}$$

$$Xs = FFTShift(X) \tag{2}$$

A coordinate grid (U_{mn}, V_{mn}) is generated with its origin at the image center. Create an ideal high-pass mask with radius $r = 8$:

$$mask2d(m,n) = \begin{cases} 1, U_{mn}^2 + V_{mn}^2 \geq r^2 \\ 0, U_{mn}^2 + V_{mn}^2 < r^2 \end{cases} \tag{3}$$

Finally use $mask2d$ to keep high frequencies:

$$X_{hp}^s = Xs \odot mask2d \tag{4}$$

Step 3: Inverse shift and inverse transform

After multiplying the shifted spectrum by the high-pass mask, apply the inverse frequency shift (IFFTShift), followed

by the inverse Fourier Transform (IFFT) to reconstruct the high-frequency image:

$$x_{hp} = IFFT\left(IFFTShift(X_{hp}^{s})\right) \quad (5)$$

Next, compute the magnitude of this complex tensor using the formula:

$$mag(x_{hp}) = |Re(x_{hp}) + iIM(x_{hp})| \quad (6)$$

Finally, compress the dynamic range with a logarithmic transform:

$$hp_{log} = \ln\left(1 + mag(x_{hp})\right) \quad (7)$$

Step 4: Normalization

Normalize the log-domain high-frequency image to [0,1] using the formula:

$$hp_{gray} = \frac{hp_{log} - m_{min}}{m_{max} - m_{min} + \varepsilon}; \varepsilon = 10^{-8} \quad (8)$$

Replicate $hp_{gray} \in \mathbb{R}^{B\times1\times224\times224}$ 3 times. Each channel is then applied the per-channel ImageNet normalization. The final result is a tensor $x_{fft} \in \mathbb{R}^{B\times3\times224\times224}$ ready to be fed into the FFT branch.

3.3 CNN backbone with VGG16

In this module, we employ the VGG16 architecture with weights pre-trained on the ImageNet dataset [11] and retain only the first seven convolutional layers. Truncating the network helps reduce model size and computational cost while still achieving good performance, as confirmed by experimental results. Both branches are initialized from the same ImageNet [11] checkpoint but maintain independent weights (transfer learning), reducing parameters and compute relative to the full model while retaining good accuracy. The output of this module is a feature tensor of shape $(B, C = 256, H = 28, W = 28)$.

3.4 Multi-head cross-attention module

To enable the network to learn the mutual relationships between features from the original image and high-frequency information, we construct a multi-head cross-attention module. This module allows each unit in the RGB image to "attend" to corresponding or related positions in the high-frequency filtered image, and vice versa, thereby establishing semantic connections and fusing information from both sources. For compatibility with the cross-attention mechanism, feature tensors from each branch are flattened and permuted from shape $B \times C \times H \times W$ to $(HW) \times B \times C$. The number of attention heads is set to: $n_{heads} = 16$ and $d_k = d_v = \frac{C}{n_{heads}}$. Cross-attention is performed bidirectionally, meaning:

(i). $RGB \leftarrow FFT$:

$$Attention_{RGB\leftarrow FFT} = softmax\left(\frac{Q_{RGB}K_{FFT}^{T}}{\sqrt{d_k}}\right)V_{FFT} \quad (9)$$

(ii). $FFT \leftarrow RGB$:

$$Attention_{FFT\leftarrow RGB} = softmax\left(\frac{Q_{FFT}K_{RGB}^{T}}{\sqrt{d_k}}\right)V_{RGB} \quad (10)$$

where:

$$Attention_{RGB\leftarrow FFT}, Attention_{FFT\leftarrow RGB} \in \mathbb{R}^{(HW)\times B\times d_v}$$

The outputs from each attention direction are combined through a linear layer to obtain tensors $Z_{RGB}, Z_{FFT} \in \mathbb{R}^{(HW)\times B\times C}$. To reduce the risk of losing important features during training, a residual connection is applied:

$$Z_{RGB} = Z_{RGB} + F_{RGB}^{flat} + \text{và } Z_{FFT} = Z_{FFT} + F_{FFT}^{flat} \quad (11)$$

Next, the tensors are permuted back to the format $B \times C \times (HW)$, and global average pooling is applied to produce a tensor of shape $B \times C$ for each branch. Finally, the two tensors are concatenated to obtain a unified representation:

$$Z = \text{Concat}(Z_{RGB}, Z_{FFT}) \in \mathbb{R}^{B\times2C} \quad (12)$$

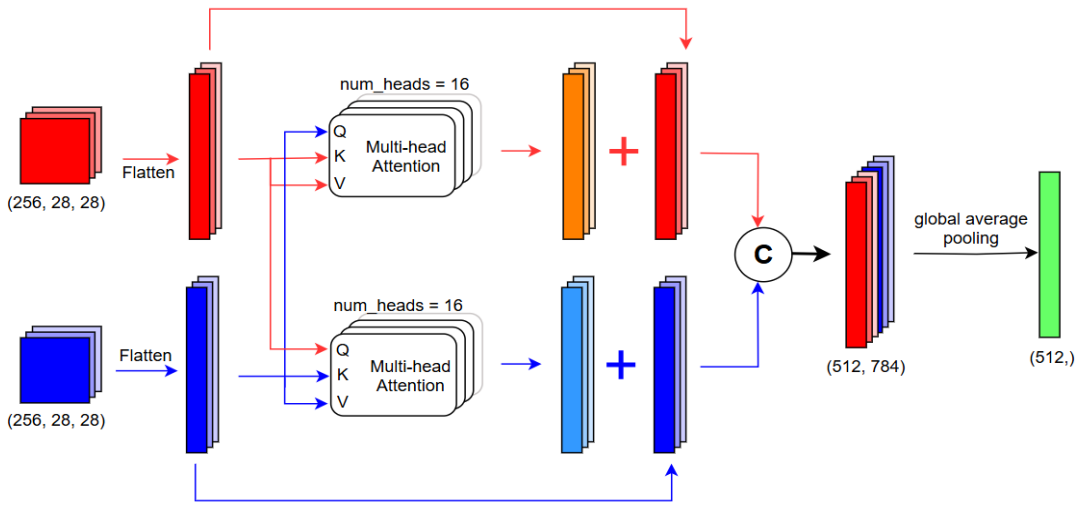

Figure 4. Proposed architecture of the VGG16 + FFT model

This module enables the model to effectively learn the relationship between spatial and frequency-domain information, enhancing its discriminative capability for the classification task. The overall process is illustrated in Figure 4.

3.5 Classification module

After fusing the features from the main and FFT branches, the resulting tensor Z with shape $(B, 2C)$, where $C = 256$, represents the feature dimension from each branch. This tensor is then passed into a classification module consisting of three fully connected layers, which are responsible for further filtering and separating the fused features to support the final classification task. The module outputs one of two predicted labels: "genuine" or "spoof".

4. MODEL TRAINING AND EVALUATION RESULTS

4.1 Dataset

The DLC-2021 [8] dataset provides a collection of genuine and spoofed images of identity documents, including national ID cards and passports from various countries. It is designed to support research and development of spoofing detection methods for these types of documents. In this study, two subsets are used: (i) genuine images classified as "or", and (ii) spoofed images captured from digital screens entitled "re". These are referred to as "genuine" and "spoof", respectively, samples for training and evaluation purposes in this paper.

The dataset follows the DLC-2021 [8] protocol and is split into train/validation/test. Training and validation use 16,264 genuine images from "MIDV-2020/clips" folder and spoofed images from "DLC-2021/re" limited to Spanish IDs, Latvian passports, and Russian internal passports; an 80/20 split is applied. The test set contains 11,006 spoof (positive) and 16,264 genuine (negative) images drawn from the remaining DLC-2021 [8] document types, fixed by the official lists "*screen_positive_test.lst*" and "*screen_negative_test.lst"*. This setup enforces cross-type generalization and ensures reproducible, baseline-comparable evaluation. Representative examples from the dataset are shown in Figure 5 and Figure 6 respectively.

Figure 5. Sample images labeled as "genuine"

Figure 6. Sample images labeled as "spoof"

4.2 Loss function and optimization process

After passing through the classification module, the model outputs a tensor with two elements corresponding to the predicted probabilities for the classes "genuine" and "spoof", denoted as $z = (z_0, z_1)$ and the true label $y \in \{0,1\}$. The loss function used is binary cross-entropy, defined as:

$$L = \frac{1}{B}\sum_{n=1}^{B} CELoss(z_n, y_n) = -\log\left(p_{y_n}\right) \tag{13}$$

The loss is minimized using the Adam optimizer, which combines the advantages of RMSProp and SGD. This allows the model to effectively learn complex features from both branches, ensuring fast convergence without overfitting the training data. The learning rate is set to $\alpha = 10^{-4}$.

4.3 Model training

During training, the batch size was set to $B = 128$. In each epoch, the data was loaded in parallel using 7 worker threads to speed up reading and preprocessing. A random shuffle was also applied before batching to ensure variety in the sample distribution. The training was performed on Google Colab using an L4 GPU (24 GB GDDR6 memory, fully supporting CUDA 11.x+, cuDNN, and TensorRT) [28]. With CUDA drivers and PyTorch Lightning integration, the GPU was accessed directly through the torch.cuda API during training. The model, which combines VGG16 and a frequency-processing branch (FFT Branch), was configured to run for a maximum of 10 epochs and early stopping callback, monitoring validation loss with mode = min, patience = 3 epochs, and min_delta = 10^{-4}. Early stopping was triggered after 7 epochs, as the validation loss no longer improved significantly, with validation loss = 0.2565. Training consumed ~20.5 GB VRAM.

4.4 Model evaluation

To ensure a fair and accurate assessment, the evaluation protocol mirrors the DLC-2021 [8] baseline: the same datasets, split definitions (official "*screen_positive_test.lst"* and "*screen_negative_test.lst"*), label convention (positive = spoof, negative = genuine). Results are reported on the fixed test split to guarantee reproducibility and comparability with prior work. Three standard metrics are reported: precision, recall, and accuracy. Precision is the proportion of predicted spoofs that are truly spoof. Recall is the proportion of actual spoofs correctly detected. Accuracy is overall rate of correct decisions. Let TP, FP, TN, FN denote true positives, false positives, true negatives, and false negatives (with "positive" = spoof). Then:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

To assess the model's accuracy, common spoof detection metrics used in this study are: Attack Presentation Classification Error Rate (APCER) [29] is the percentage of spoofed samples that are incorrectly classified as genuine. And Bona Fide Presentation Classification Error Rate (BPCER) [29] is the percentage of genuine samples that are incorrectly classified as spoofed. Half Total Error Rate (HTER) [30, 31] is calculated as the average of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR). In spoof detection problems, it is typically assumed that: $FAR = APCER;\ FRR = BPCER$. Therefore, HTER can be obtained as shown in Eq. (17):

$$HTER = \frac{PCER + BPCER}{2} \tag{17}$$

These three metrics provide a balanced evaluation between the ability to detect attacks and the ability to correctly recognize genuine samples. They follow ISO standards for electronic identification and biometrics [29]. To measure performance, the model is evaluated from several aspects, including computational complexity (FLOPs), number of parameters (Params), average inference speed (measured in seconds per image).

For comparison, two reference models were evaluated alongside the proposed model:

(i) DLC-2021 baseline [8]

(ii) Boundary moiré detector [13]

The evaluation results of all models are summarized in Table 1.

Table 1. Evaluation results of accuracy of DLC-2021 baseline, Boundary moiré detector and VGG16+FFT

Model	Precision (%)	Recall (%)	Accuracy (%)
DLC-2021 baseline	85.89	89.03	89.67
Boundary moiré detector	93.5	90.2	92.1
VGG16+FFT	93.64	88.90	91.68

Regarding spoof detection accuracy, with the DLC-2021 protocol and identical test lists, the Boundary moiré detector attains the highest accuracy (92.1%) with precision 93.5% and recall 90.2%. The proposed VGG16+FFT is close behind (accuracy 91.68%, precision 93.64%, recall 88.90%), while the DLC-2021 baseline yields accuracy 89.67% with precision 85.89% and recall 89.03% (Table 1).

Table 2. Evaluation results of average inference speed on Intel Xeon @ 2.20 GHz (x86-64), 6-core / 12-thread

Model	Folder name: grc_passport/01.or 0004	Folder name: grc_passport/01. re0001
DLC-2021 baseline	0.147	0.151
Boundary moiré detector	-	-
VGG16+FFT	0.120	0.120

Regarding Inference Speed, on an Intel Xeon @ 2.20 GHz (x86-64), 6-core/12-thread CPU, VGG16+FFT processes representative samples in 0.120 s/image for both original and recaptured cases, outperforming the DLC-2021 baseline (0.147-0.151 s/image on the same files). Runtime for the Boundary moiré detector is not available under this setup (Table 2).

VGG16 + FFT has a complexity of about 40.95 GFLOPs, mainly because of its two-branch design combined with the cross-attention module that requires intensive numerical transformations. About the number of parameters (Params): VGG16+FFT uses only about 4.29 million parameters, which is much fewer than the original VGG16. VGG16 + FFT achieves a low BPCER (genuine samples misclassified as spoofed) of just 5.33% and APCER (spoofed samples misclassified as genuine) at 14.15%, indicating high sensitivity to spoofed cues. The overall HTER is 9.74%.

Summary, within the same evaluation protocol, the Boundary moiré detector delivers the top accuracy, while VGG16+FFT offers a favorable accuracy-speed trade-off and is faster than the DLC-2021 baseline on the measured samples. These results indicate that the proposed model is suitable for deployment scenarios that require both reliability and efficient CPU-side inference.

4.5 Ablation study

All ablations follow the same DLC-2021 [8] protocol (positive = spoof), fixed test lists, and 224×224 center-crop preprocessing. Metrics are precision, recall, and accuracy.

Let VGG16+FFT (proposed) with precision 93.64%, recall 88.90%, accuracy 91.68% serve as the reference.

The following investigations have been carried out:

•Replace CrossAttentionFusion with ConcatGAP: A simple concatenation followed by global average pooling is used to merge the spatial and FFT streams.

•Replace VGG16+FFT with two channels of VGG16: The FFT branch is removed; two spatial (VGG16) streams are provided instead.

•Use VGG16-only (single branch): The VGG16+FFT and the CrossAttentionFusion module are removed, leaving a single VGG16 spatial branch followed by the same classification head as in the full model. All training and evaluation settings are kept identical. Cross-attention is not applicable in the single-branch setting.

Table 3. Evaluation results of ablation studies

Model	Precision (%)	Recall (%)	Accuracy (%)
Replace CrossAttentionFusion with ConcatGAP	91.29	83.44	88.10
Replace VGG16+FFT with two channels of VGG16	91.38	84.80	88.73
Use VGG16-only (single branch)	87.67	86.35	87.48

Simple GAP reduces precision to 91.29%, recall to 83.44%, and accuracy to 88.10%. The loss of cross-modal alignment increases missed spoofs (FN), indicating that cross-attention is important for merging complementary spatial-frequency cues.

Duplicating the spatial stream (no FFT) yields 91.38% precision, 84.80% recall, and 88.73% accuracy. Without frequency-domain evidence, the model under-detects moiré artifacts, confirming that FFT features are not redundant with spatial features (Table 3).

Removing the VGG16+FFT and cross attention fusion entirely lowers recall and overall accuracy relative to the reference, as complementary frequency cues and any alignment mechanism are absent. Cross-attention is invalid in

the single-branch architecture, so this comparison isolates the contribution of the FFT pathway and highlights the cost of discarding it.

5. DISCUSSION

The experimental results demonstrate that integrating a VGG16-based CNN with Fourier-domain analysis effectively achieves the study's primary goal: building a spoof detection model that is both accurate and parameter-efficient. The proposed dual-branch design, with one branch extracting spatial characteristics and the other highlighting frequency-enhanced features, reached an accuracy of 91.68%. This represents a clear improvement over baseline VGG16 and other benchmark models, confirming that frequency information provides critical advantages in detecting subtle spoofing artifacts often missed in spatial-only analysis. This design with cross-attention fusion allowed the system to exploit complementary spatial-frequency interactions and achieve state-of-the-art accuracy.

At the same time, the findings highlight an important trade-off: although the FFT branch adds computational overhead, its contribution to accuracy makes the extra cost worthwhile. Model sensitivity to design choices, such as the FFT mask radius (set to $r = 8$ in this study) and the cross-attention configuration, also indicates the need for systematic parameter optimization. Future research should therefore focus on detailed parametric studies and broader experimentation to refine these components and further balance the trade-off between accuracy, efficiency, and computational cost.

6. CONCLUSIONS

This paper introduced a dual-branch CNN architecture for ID-card spoof detection that integrates spatial and frequency-domain analysis within a parameter-efficient design. The main branch employed a truncated VGG16 backbone to reduce redundancy, while the FFT branch captured high-frequency artifacts characteristic of spoofing. By fusing these complementary features through a multi-head cross-attention mechanism, the model achieved an accuracy of 91.68%, outperforming baselines and demonstrating that frequency characteristics significantly enhance the detection of subtle spoofing traces.

Under the DLC-2021 protocol, the proposed VGG16+FFT achieves nearly the same performance as the Boundary moiré detector (91.68% vs. 92.1% accuracy), while clearly outperforming the DLC-2021 baseline model (89.67% accuracy). This approach offers faster inference (0.120 s/image) than the DLC-2021 baseline (0.147-0.151 s/image). The ablation analysis confirms the complementary roles of the dual-branch design for robust spoof detection.

A key contribution of this study is showing that high accuracy can be achieved without relying on a heavy network. The design maintains a compact model size while still benefiting from the FFT branch. Although this branch increases computational complexity, the substantial accuracy gain highlights a meaningful trade-off—making the approach highly relevant for security-critical applications such as eKYC systems.

Performance may degrade on high-resolution or high-refresh-rate displays (e.g., OLED/120 Hz) whose moiré patterns differ from the training distribution; the FFT emphasis on high frequencies can amplify such device-specific signatures. Complex or cluttered backgrounds and imperfect ID localization may also confound the frequency branch, especially when background textures contain periodic structures similar to screen artifacts. Our evaluation is limited to DLC-2021; cross-database generalization remains untested. Finally, although the model is compact on a desktop-class GPU/CPU, embedded devices face tighter budgets for compute, memory, and energy; real-time throughput and latency can drop without additional optimization.

Future work should focus on systematic parameter optimization, particularly refining FFT configurations and attention mechanisms, as well as testing lightweight backbones beyond VGG16 to further improve efficiency. Expanding the dataset to include more diverse forgery types (e.g., printed, laminated, or digitally altered IDs) and using other datasets for evaluation. Together, these directions offer a clear pathway toward building more accurate, efficient, and resilient identity verification systems.

REFERENCES

[1] Lan, H., Van, H. (2023). Có 40 ngân hàng chính thức triển khai quy trình mở tài khoản thanh toán eKYC. https://vneconomy.vn/co-40-ngan-hang-chinh-thuc-trien-khai-quy-trinh-mo-tai-khoan-thanh-toan-ekyc.htm.

[2] Nhân Dân. Cashless payment on the rise in Vietnam. https://en.nhandan.vn/cashless-payment-on-the-rise-in-vietnam-post127059.html.

[3] Vinh, P. (2024). Người dân bị lừa đảo khoảng 10.000 tỷ đồng trên không gian mạng. https://vneconomy.vn/nguoi-dan-bi-lua-dao-khoang-10-000-ty-dong-tren-khong-gian-mang.htm.

[4] VietnamPlus. Vietnamese lose nearly 737 USD to each online scam in 2023. https://en.vietnamplus.vn/vietnamese-lose-nearly-737-usd-to-each-online-scam-in-2023-post276427.vnp.

[5] Gonzalez, R.C., Woods, R.E. (2002). Chapter 4: Filtering in frequency domain. In Digital Image Processing, pp. 199-310.

[6] Jain, A. (1989). Fundamentals of Digital Image Processing. Prentice-Hall Inc., Englewood Cliffs, NJ.

[7] Schatzman, J. (1996). Accuracy of the discrete fourier transform and the fast Fourier transform. SIAM Journal on Scientific Computing, 17(5): 1150-1166. https://doi.org/10.1137/S1064827593247023

[8] Polevoy, D.V., Sigareva, I.V., Ershova, D.M., Arlazarov, V.V., Nikolaev, D.P., Ming, Z., Luqman, M.M., Burie, J.C. (2022). Document liveness challenge dataset (DLC-2021). Journal of Imaging, 8(11): 181. https://doi.org/10.3390/jimaging8110289

[9] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[10] Koonce, B. (2021). ResNet 50. In Convolutional Neural Networks with Swift for TensorFlow: Image Recognition and Dataset Categorization. Apress, Berkeley, CA, pp. 63-72.

[11] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2009). ImageNet: A large-scale hierarchical image

database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[12] Bulatov, K., Emelianova, E., Tropin, D., Skoryukina, N., Chernyshova, Y., Sheshkus, A., Usilin, S., Ming, Z., Burie, J.C., Luqman, M.M., Arlazarov, V.V. (2022). MIDV-2020: A comprehensive benchmark dataset for identity document analysis. Computer Optics, 46(2): 252-270. https://doi.org/10.18287/2412-6179-CO-1006

[13] Kunina, I.A., Sher, A.V., Nikolaev, D.P. (2023). Screen recapture detection based on color-texture analysis of document boundary regions. Computer Optics, 47(4): 650-657. https://doi.org/10.18287/2412-6179-CO-1237

[14] Markham, R.P., López, J.M.E., Nieto-Hidalgo, M., Tapia, J.E. (2024). Open-Set: ID card presentation attack detection using neural style transfer. IEEE Access, 12: 68573-68585. https://doi.org/10.1109/ACCESS.2024.3397190

[15] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[16] Al-Ghadi, M., Voerman, J., Coustaty, M., Lessard, O., Sidere, N. (2025). IDTrust: Deep identity document quality detection with bandpass filtering. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Tucson, AZ, USA, pp. 668-675. https://doi.org/10.1109/WACVW65960.2025.00081

[17] Boned, C., Talarmain, M., Ghanmi, N., Chiron, G., Biswas, S., Awal, A.M., Ramos Terrades, O. (2024). Synthetic dataset of id and travel documents. Scientific Data, 11(1): 1356. https://doi.org/10.1038/s41597-024-04160-9

[18] Guan, H., Wang, Y., Xie, L., Nag, S. et al. (2024). IDNet: A novel dataset for identity document analysis and fraud detection. arXiv preprint arXiv:2408.01690. https://doi.org/10.48550/arXiv.2408.01690

[19] Zhang, C.J., Gill, A.Q., Liu, B., Anwar, M.J. (2025). AI-based identity fraud detection: A systematic review. arXiv preprint arXiv:2501.09239. https://doi.org/10.48550/arXiv.2501.09239

[20] Zeng, D., Gao, L., Fang, H., Xiang, G., Feng, Y., Lu, Q. (2023). Bandpass filter based dual-stream network for face anti-spoofing. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, pp. 6403-6410. https://doi.org/10.1109/CVPRW59228.2023.00681

[21] Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J. (2022). EfficientFormer: Vision transformers at MobileNet speed. In Proceedings of the 36th International Conference on Neural Information Processing Systems, ew Orleans, LA, USA, pp. 12934-12949.

[22] Li, Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y., Tulyakov, S., Ren, J. (2023). Rethinking vision transformers for MobileNet size and speed. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, pp. 16843-16854. https://doi.org/10.1109/ICCV.2023.01553

[23] Yu, Z., Zhao, C., Wang, Z., Qin, Y., et al. (2020). Searching central difference convolutional networks for face anti-spoofing. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 5294-5304. https://doi.org/10.1109/CVPR42600.2020.00534

[24] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, pp. 612-618. https://doi.org/10.1109/FG.2017.77

[25] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z. (2012). A face antispoofing database with diverse attacks. In 2012 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, pp. 26-31. https://doi.org/10.1109/ICB.2012.6199754

[26] Shinde, P., Raundale, A.R. (2024). Face and liveness detection with criminal identification using machine learning and image processing techniques for security system. IAES International Journal of Artificial Intelligence (IJ-AI), 13(1): 722-729. https://doi.org/10.11591/ijai.v13i1.pp722-729

[27] Nixon, M.S., Aguado, A.S. (2008). Feature Extraction and Image Processing. Academic Press, London.

[28] Google Cloud Platform. (2024). GPUs on Compute Engine - GPU machine types. Google. https://cloud.google.com/compute/docs/gpus.

[29] ISO/IEC 30107-3. (2023). Information technology – Biometric presentation attack detection – Part 3: Testing and reporting. International Organization for Standardization. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:30107:-3:ed-2:v1:en.

[30] Bengio, S., Mariéthoz, J. (2004). A statistical significance test for person authentication. https://core.ac.uk/reader/147915360.

[31] Luevano, L.S., Martínez-Díaz, Y., Méndez-Vázquez, H., González-Mendoza, M., Frey, D. (2024). Assessing the performance of efficient face anti-spoofing detection against physical and digital presentation attacks. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, pp. 1021-1028. https://doi.org/10.1109/CVPRW63382.2024.00108