electronics

Article

PBNet: Combining Transformer and CNN in Passport Background Texture Printing Image Classification

Jiafeng Xu [1,*], Dawei Jia [1], Zhizhe Lin [2,*], Teng Zhou [2,3], Jie Wu [4] and Lin Tang [5]

[1] School of Immigration Management, China People's Police University, Guangzhou 510663, China; jiadawei@cppu.edu.cn
[2] School of Cyberspace Security, Hainan University, Haikou 570228, China; teng.zhou@hainanu.edu.cn
[3] Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou 324003, China
[4] School of Wu Shu, Henan University, Kaifeng 610054, China; wujie@hrbipe.edu.cn
[5] School of Electronic Information and Electrical Engineering, Chengdu University, Chengdu 610106, China; tanglin@cdu.edu.cn
* Correspondence: xujiafeng@cppu.edu.cn (J.X.); l.zhizhe@gmail.com (Z.L.)

Abstract: Passport background texture classification has always been an important task in border checks. Current manual methods struggle to achieve satisfactory results in terms of consistency and stability for weakly textured background images. For this reason, this study designs and develops a CNN and Transformer complementary network (PBNet) for passport background texture image classification. We first design two encoders by Transformer and CNN to produce complementary features in the Transformer and CNN domains, respectively. Then, we cross-wisely concatenate these complementary features to propose a feature enhancement module (FEM) for effectively blending them. In addition, we introduce focal loss to relieve the overfitting problem caused by data imbalance. Experimental results show that our PBNet significantly surpasses the state-of-the-art image segmentation models based on CNNs, Transformers, and even Transformer and CNN combined models designed for passport background texture image classification.

Keywords: passport; background texture; Convolutional Neural Network; Transformer

check for updates

Citation: Xu, J.; Jia, D.; Lin, Z.; Zhou, T.; Wu, J.; Tang, L. PBNet: Combining Transformer and CNN in Passport Background Texture Printing Image Classification. *Electronics* **2024**, *13*, 4160. https://doi.org/10.3390/electronics13214160

Academic Editor: Hyunjin Park

Received: 24 September 2024
Revised: 15 October 2024
Accepted: 21 October 2024
Published: 23 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The security and authenticity of passports, as important credentials for cross-border identity verification, play a crucial role in national security. However, the crime of passport forgery is closely related to a wide range of cross-border criminal activities [1,2]. Document fraud not only poses a great challenge to border management but also poses a serious threat to global security [3].

The quest for automated methods to identify and evaluate security documents has become increasingly vital for achieving a more rapid, cost-effective, and error-minimized assessment process [4]. Computer vision technologies have emerged as a valuable asset either supplementing or even supplanting the need for human analysis in identifying counterfeit characteristics [5]. Researchers have proposed a variety of algorithms to ascertain the authenticity of documents [6,7]. Nevertheless, the challenge of counterfeit detection continues to be a domain of ongoing research. Many existing solutions still necessitate the use of specialized equipment for capturing the images that will subsequently be scrutinized by these algorithms. In the pursuit of advancements, numerous researchers have ventured into employing sophisticated algorithms such as Convolutional Neural Networks (CNNs) and Transformer models to enhance the detection process [8,9]. These cutting-edge approaches have demonstrated promising results, contributing to the progress in the field of document security analysis. Although existing research has attempted to process images for recognition using different computer vision methods, there are still some challenges in practical applications.

1.1. Images with Different Local Detail Features and Decentralized Global Features

As shown in Figure 1, the passport background texture image contains a variety of local details and global features, such as variations in the width of the lines, differences in the light and darkness of the texture, and the complexity of the pattern, which constitute key elements of the passport's security features. Figure 2 shows four different passport background textures. Lithography printing relies on the principle that oil and water do not mix [10]. Lithography printing shows solid lines, which are smooth and flat, with uniform colors and no grid pattern. Currently, most passport background textures are printed with lithography. Transformed design is a printing technique that utilizes the variation in the depth or direction of lines on a document plate to print unconventional text or patterns. It forms three-dimensional text or patterns by altering the direction of the lines. Due to the requirement for professional cutting-edge printing equipment to reproduce lithographic printing and transformed design, counterfeit patterns on passports are typically created using laser printing or inkjet printing. The patterns from laser printing are composed of dots with small ink dots that do not smudge and are arranged in a more regular pattern. Different printer models display different dot matrix patterns. Patterns from inkjet printing, when observed under magnification, have larger ink dots that are arranged haphazardly and without a discernible pattern. The ink dots spread and permeate into the paper fibers [11].

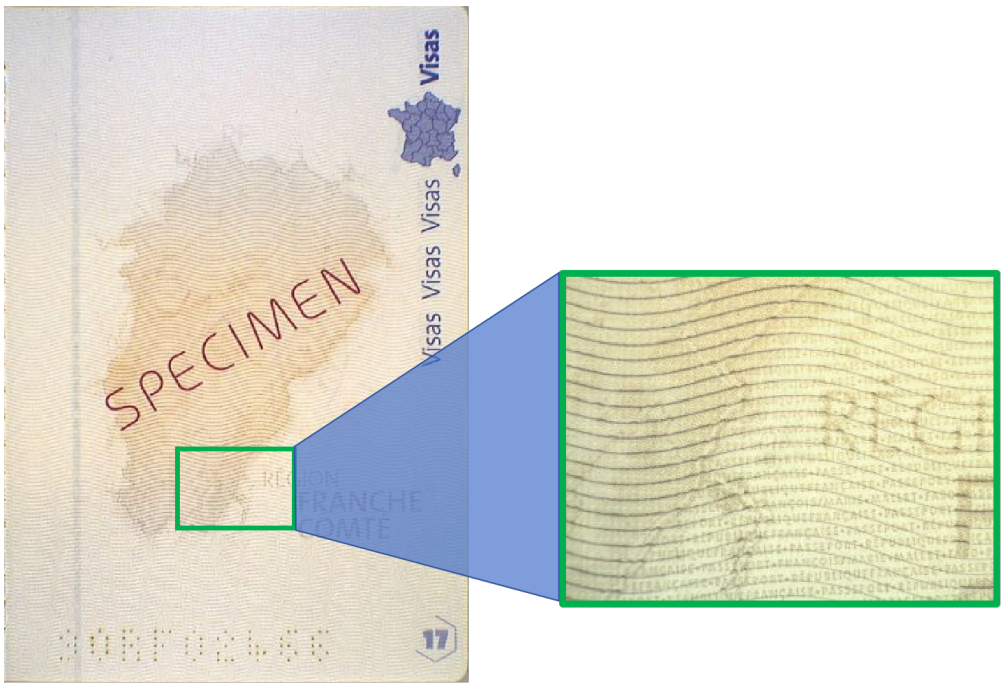


Figure 1. The passport background texture image contains a variety of local details and global features.

The passport background texture is designed with fine lines and low-contrast color tones, which are not easily recognized by manual inspection and pose a challenge to machine recognition techniques. Convolutional Neural Network (CNN) algorithms may encounter difficulties in processing such images as CNNs are usually good at capturing local features and have limited ability to capture global features [12]. On the other hand, Transformer models, while excellent at handling global dependencies, may not be sensitive enough to subtle local features [13]. Therefore, a single CNN or Transformer model makes it difficult to perform the task of recognizing and classifying passport background texture images.

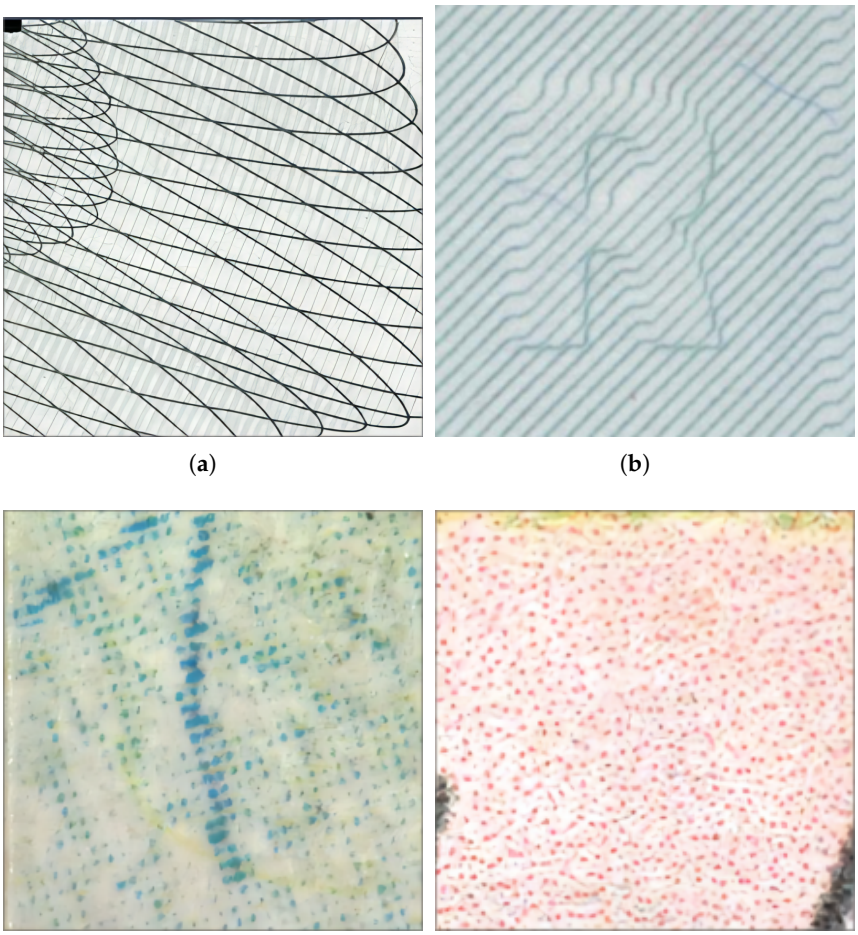


Figure 2. Confusion matrices of four models in the test set. (**a**) Lithography printing. (**b**) Transformed design. (**c**) Laser printing. (**d**) Inkjet printing.

1.2. Existing Models Are Difficult to Deploy on Mobile Devices

In deep learning image recognition and classification tasks, while efficient models such as AlexNet, VGG, Inception series, ResNet, and DenseNet continue to emerge, their number of parameters is unusually large. For example, AlexNet [14] contains about 62 million parameters, while VGG [15] contains 130 to 140 million, and ResNet-50 contains [16] about 25 million. These models are difficult to apply on mobile devices with limited storage, computation, and battery due to their computational intensity and large number of parameters. Although lightweight models such as MobileNet [17] and SqueezeNet [18] enhance the applicability on mobile devices by reducing parameters and computational requirements, their accuracy on image underline recognition still needs to be improved. How to achieve model lightweighting while maintaining accuracy is an urgent research topic.

1.3. Sample Imbalance

The difficulty in obtaining forged passport images, especially those produced using laser or inkjet printing, caused the number of forged images in the training data to be much lower than the number of real images, resulting in an imbalance in labeling frequency in the dataset. It causes the model to favor the majority class and ignore the minority class, reducing the ability to recognize rare types of forgeries. Therefore, specific algorithms are needed to enhance the model's ability to generalize to all classes and improve the recognition performance for minority classes.

In this paper, we propose a model termed PBNet to alleviate the above problem. For the first problem, we utilize a parallel structure that allows information interaction between the CNN and the Transformer, where the CNN can provide the Transformer with enriched local features and the Transformer can guide the CNN to more efficient feature extraction through its top–down attention mechanism, thus improving the recognition of passport background texture images. For the second problem, we adopt a lightweight CNN model, MobileNet v2, and build a parallel structure with Transformer. MobileNet v2 significantly reduces the number of parameters and computational complexity of the model by using Depthwise Separable Convolution, while maintaining a high accuracy, making it ideal for deployment on mobile devices. For the third problem, we introduce a loss function, focal loss, which is expected to increase the accuracy of the model by focusing more on minority and hard-to-classify samples.

The contributions of this work are as follows:

- Passport Background Texture Dataset: We constructed a passport background texture image dataset that includes four types of passport background textures from various countries.
- Parallel Hybrid Architecture for Passport Background Texture Image Classification: We introduce a new parallel hybrid architecture that combines CNN and Transformer. The CNN delivers comprehensive local features to the Transformer through its bottom–up extraction process, while the Transformer, in turn, enhances the CNN's feature extraction capabilities with its top–down attention mechanisms. This information interaction strengthens the collaboration between the two components, improving the model's overall performance. Concurrently, this interaction allows the CNN, which is good at extracting local features, to better model the global features extracted by the Transformer.
- Cross-Branch Interaction via Feature Enhancement Module: To improve the model's nonlinear representation and feature representation capabilities in image classification tasks, we introduce a feature enhancement module, which facilitates cross-branch communication within the model and helps to improve the ability to differentiate between different classes of image features, thus enhancing classification accuracy.
- Comprehensive Evaluation and Performance: We evaluated our proposed framework PBNet on a self-constructed dataset and compared it with other models. The experimental results show that PBNet displays superior performance, highlighting the effectiveness of our model in the task of passport background texture image classification.

The remainder of this paper is organized as follows. We first discuss related works in Section 2. Section 3 describes the research process and the details of the proposed model. In Section 4, the performance analysis is discussed. Finally, a conclusion with a few oncoming work purposes is explained in Section 5.

2. Related Work

2.1. CNN-Based Image Classification Method

In recent years, deep learning-based methods, especially CNN-based methods, have achieved significant improvements in the image classification problem. The Fully Convolutional Network (FCN) proposed by Long et al. [19] has become one of the mainstream techniques in the field of semantic segmentation. It inspired subsequent researchers to adopt deeper and more broadly connected network structures or more efficient designs, such as VGG, GoogleNet, and ResNet. Burri et al. [20] explored the effectiveness of transfer learning in image classification tasks by optimizing a CNN model, especially after pretraining the model on the ImageNet dataset, and achieved significant results in medical image analysis applications such as COVID-19 detection and Alzheimer's disease classification. Furthermore, Han et al. [21] proposed a novel dynamic multi-scale CNN (DM-CNN), which improves the accuracy and reliability of medical image classification through uncertainty quantization. The design of DM-CNN takes into account the complexity of medical images by means of a dynamic multi-scale feature fusion module (DMFF), hierarchical

dynamic uncertainty quantization attention (HDUQ- Attention), and multi-scale fusion pooling method (MF Pooling), which effectively improves the model's generalization ability and assessment of uncertainty. Zhang et al. [22] introduced a method for detecting surface defects on winter jujubes using an improved AlexNet. The improved model achieved the highest accuracy in the validation set during comparative experiments. TS-CNN performs well in the task of multi-region medical image classification, especially in dealing with images with scattered and randomly shaped lesions. Jiang et al. [23] proposed a CNN pruning method based on multi-objective feature map selection (MOP-FMS). This method not only focuses on the accuracy of the network but also takes the number of floating point operations (FLOPs) as the pruning target, which effectively reduces the computational cost while maintaining the accuracy through the feature map selection coding method and a specific genetic algorithm. Finally, Zhang et al. [24] proposed a model called TMOE-CNN for hyperspectral image classification. TMOE-CNN achieves a significant improvement in computational efficiency while maintaining a high classification accuracy through the design of specific individual representations, a tree-shaped multi-branching super-network structure, and efficient crossover and mutation operators.

In order to satisfy the stringent requirements of computational resources and memory occupation in practice, a number of lightweight models have emerged. These models, such as MobileNet, SqueezeNet, and ShuffleNet, aim to significantly reduce the computation and memory requirements while maintaining high performance to adapt to different application scenarios and computing environments. Pandiri et al. [25] developed a lightweight CNN (Light-SoilNet) for fast and cost-effective soil image classification. Li et al. [26] proposed an innovative lightweight Convolutional Spiking Neural Network (CSNN) method for fire detection based on acoustics.

2.2. Transformer-Based Image Classification Method

Despite the excellent performance of CNN-based methods, most of them are not practical in the real world due to heavy computational complexity. To solve this problem, Transformer was introduced for machine vision tasks.

With its powerful self-attention mechanism, the Transformer model has achieved great success in Natural Language Processing tasks, which has motivated researchers to apply it to computer vision tasks. The Transformer model brings a new perspective to image analysis by capturing long-range dependencies in the input data. Ran et al. [27] proposed a framework that fuses depth Transformer and few-sample learning (DT-FSL) to effectively capture the nonlocal spatial–spectral information of hyperspectral images, which further improves the model's ability to recognize small-scale targets. Zhou et al. [28] proposed a Transformer-based feature learning network (FL-Tran), which solves the limitations of existing methods in small-scale object recognition and significantly improves the model performance through multi-scale feature fusion and spatial attention mechanism. Wu et al. [29] proposed a cascade neural network model constructed of a Deep Neural Network and a Transformer, which achieved high-precision predictions of multiple metrics in performance prediction and optimization. Wang et al. [30] proposed a new dual-context network, DCN-T, which effectively encodes intra- and inter-region contextual information using Transformer and integrates the segmentation results of all trispectral images through a voting mechanism to improve the classification accuracy further.

With the further study of Transformer, some researchers have started to try to combine it with CNN in order to utilize the advantages of both. Gong et al. [31] introduced a novel deepfake detection approach utilizing the Swin Transformer, which demonstrates enhanced robustness and capability to discern more complex data relationships compared to traditional supervised classification methods. Yuan et al. [32] designed a CNN and Transformer Complementary Network (CTC-Net), which enhances feature fusion through feature complementary module (FCM) and multi-level jump connections. In addition, Zhang et al. [33] developed a lightweight Transformer network for hyperspectral image

classification and optimized the training process with a controlled multi-class hierarchical (CMS) sampling strategy.

2.3. Deep Learning-Based Passport Detection

Traditional passport detection methods mainly rely on manual work, but with the continuous development of AI, the application of deep learning algorithms to the field of secure document analysis has gradually become the main research direction of researchers. For example, Al-Ghadi et al. [34] proposed a forgery detection technique based on guilloche patterns, which utilizes a twin network to capture subtle differences between forged and authentic passports. Sirajudeen et al. [35], on the other hand, focus on Azerbaijani passports by identifying possible forged areas such as text and holograms and applying Convolutional Neural Networks for classification. Gonzalez et al. [36] further developed this area by designing a hybrid two-stage architecture that achieves high-accuracy forgery detection on a Chilean passport database through deep learning. In addition, Ghanmi et al. [37] proposed a deep learning-based ID verification framework to verify authenticity by comparing the feature vectors of the reference image and the query image. Xu et al. [38] proposed a lightweight PSFNet model for passport printing anti-counterfeiting patterns, which enhances the feature learning by a pixel attention module and focus loss function. Jeny et al. [39] used ResNet50 for the passport covers for national recognition and achieved an average accuracy of 98.56%.

In our review of the existing literature, we note that there is no research on passport background texture classification, although computer vision techniques have made significant progress in several fields. In this study, we apply computer vision techniques to the automatic identification and classification of passport background texture, aiming to significantly reduce the labor intensity and resource consumption of manual review. Particular attention is paid to the design of a lightweight network for deployment on mobile devices, such as law enforcement cameras, and to maintain high accuracy even in complex passport background texture environments, thus providing law enforcement officers with an efficient and reliable auxiliary tool to support their daily work decisions.

3. Methodology

3.1. Research Process

First, we collected passport background texture printing images through public sources and built a dataset. Then, we introduced a new parallel hybrid architecture that combines CNN and Transformer. Ultimately, upon the completion of data training, we evaluated our model's efficacy.

3.2. The Details of the Proposed Model

Studies have shown that long-distance relationships between image pixels are critical for computer vision tasks. Traditional Convolutional Neural Networks have advantages in local feature extraction, but it is difficult to capture global feature information. In contrast, the Transformer model can effectively extract global feature information by utilizing the self-attention mechanism and the multi-layer perceptron structure. In order to fully utilize the respective advantages of CNN and Transformer, this paper proposes a convolutional attention fusion network PBNet. PBNet adopts a parallel structure and consists of a CNN branch and a Transformer branch, which follow the design of MobileNet v2 and the Transformer model, respectively. Due to the differences in the features extracted from the CNN and Transformer branches, they cannot be directly fused. Therefore, in this paper, we design a feature enhancement module to eliminate the differences between them and fuse the features of the two styles. The feature fusion module consists of three parts: a convolutional layer with a convolutional kernel size of (1×1), an average pooling layer, and a regularization layer. PBNet consists of a convolutional residual module, a self-attention module, a feature enhancement module, and a fully connected layer, where the feature enhancement module serves as a bridge between the convolutional residual module

and the self-attention module for eliminating the feature differences between them. The overall structure of PBNet is shown in Figure 3.

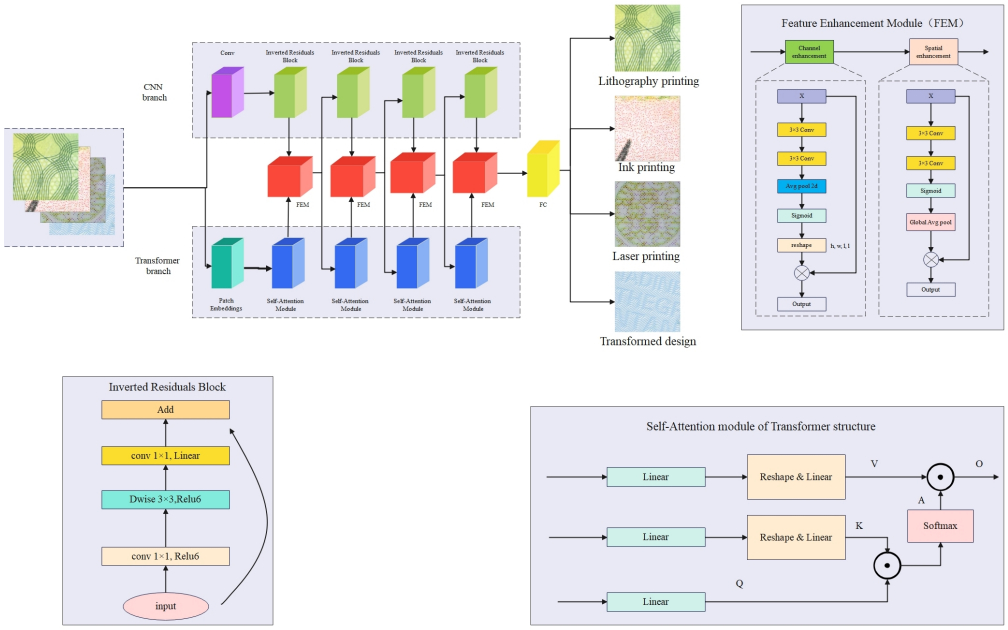


Figure 3. PBNet is a fusion of MobileNet v2 and Transformer. When the image is input into PBNet, it is processed by both the CNN branch and Transformer branch. In PBNet, the feature extraction process includes the outputs of the residual block and the self-attention module, which are simultaneously input into the feature fusion module for feature fusion. The fused feature map is then passed as input to the next residual block and self-attention module, and finally, the features output from the last layer of the feature fusion module is fed into the fully connected layer for classification.

3.2.1. Inverted Residuals Block

The inverted residuals block of the CNN branch adopts the residuals block of MobileNet v2. The residuals block has been proved in ResNet, which can help to improve the accuracy of building a deeper network, so a similar block is introduced in MobileNet v2. MobileNet v2 firstly expands the channels of feature maps through (1×1) point-by-point convolution operation to enrich the number of features and then improve the accuracy. This process is exactly the reverse order of the residual block, which is the origin of the inverse residual: (1×1)(upconvolution)–>(3×3) (dw conv + relu)–>(1×1) (dw conv + linear transformation).

3.2.2. Feature Enhancement Module

We draw on the convolutional block attention module [40] to propose our feature enhancement module, which consists of a channel enhancement module and a spatial enhancement module. The channel enhancement module and spatial enhancement module each concentrate on distinct aspects of feature response. The channel module emphasizes channel interdependencies, while the spatial module highlights relationships between spatial locations. By combining these attentions through multiplication, the model adeptly refines features, enabling superior capture and the representation of image characteristics across both channel and spatial dimensions.

Given an input $X \in \mathbb{R}^{C \times H \times W}$, FEM sequentially infers a 1D channel enhancement map $M_c \in \mathbb{R}^{C \times 1 \times 1}$ and a 2D spatial enhancement map $M_s \in \mathbb{R}^{1 \times H \times W}$. The overall process can be summarized as shown below:

$$X' = M_c(X) \otimes X, X'' = M_s(X) \otimes X', \quad (1)$$

The channel enhancement module is used to adaptively tune the feature responses of different channels to improve sensitivity to different features. It also allocates resources on each convolutional channel with the unidimensional tuning of the z-axis.

In addition, the channel enhancement module enhances the modeling of different features by learning the weights of each channel, enabling the model to adaptively adjust the importance of channel features. First, the module receives input X and then passes through two 3×3 convolutional layers to extract spatial features at different scales. This is followed by an average pooling layer in order to reduce the spatial dimensionality of the features and retain important information. After average pooling, each feature map is passed through a sigmoid activation function that compresses the values between 0 and 1 to generate a weight map. The result of average pooling is reshaped into h, w, l, and l. Finally, the reshaped data are used as an output for subsequent network layers or as a final feature representation.

In short, the channel enhancement module is computed as follows:

$$M_c(x) = \sigma(\text{AvgPool}(\text{Conv}_2(\text{Conv}_1(x))) \tag{2}$$

$Conv_1$, $Conv_2$ denote convolutional layers and $AvgPool$ means average pooling. The sigmoid activation function, denoted by σ, is a mathematical function that maps any real-valued number to a value between 0 and 1. It is defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

The spatial enhancement module is designed to adaptively adjust the feature response to optimize for features at different spatial locations in the feature map, thereby enhancing the model's sensitivity to these locations. This adaptive tuning helps the model to capture detailed changes in the image more accurately. In particular, the spatial attention mechanism, often referred to as the "look where" module, achieves a fine-grained tuning of the 2D space by applying the attention mechanism at each location of the feature map. Specifically, the mechanism dynamically adjusts these weights by calculating the weights of each point (x, y) in the feature map, enabling the model to allocate more computational resources and attention to those regions that contain important information. The Spatial Enhancement Module (SEM) first accepts the input feature map X and then goes through the same two 3×3 convolutional layers to extract and enhance the spatial features in the input feature map. The output of the convolutional layers is then passed through a sigmoid activation function, which compresses the output values to between 0 and 1. The output of the sigmoid activation is then passed through global average pooling, which compresses the features of each channel into a single value, captures the global spatial information, and finally outputs the result.

The spatial enhancement module is computed as follows:

$$M_s(x) = \text{GlobalAvgPool}(\sigma(\text{Conv}_2(\text{Conv}_1(x)))) \tag{4}$$

σ denotes the sigmoid function, $Conv_1$, $Conv_2$ denote convolutional layers and $GlobalAvgPool$ represents global average pooling.

Such an adjustment not only improves the model's ability to recognize key features but also enhances the model's ability to differentiate between different regions in the image, improving the overall performance in tasks such as image classification, target detection, and semantic segmentation.

Given a feature map, the feature enhancement module is able to sequentially generate attentional feature map information in both channel and spatial dimensions, and then the two types of feature map information are multiplied with the previous original input feature map for adaptive feature correction to produce the final feature map.

3.2.3. Focal Loss

Focal loss [41] is specifically designed to solve the problem of category imbalance in deep learning. It was originally proposed in 2017 and has been applied in target detection tasks in the image domain, significantly improving the performance of the model when dealing with unbalanced datasets [42,43]. Focal loss introduces a dynamic tuning factor by improving the standard cross-entropy loss function, which adjusts the loss weight of each sample according to the difficulty of categorization of the sample. Specifically, for those samples that the model can already easily and correctly classify, focal loss decreases their loss contribution, whereas for those samples that are difficult to classify, focal loss increases their loss weight, prompting the model to focus more on the learning of these samples. Through this mechanism, focal loss effectively guides the model training process and optimizes the model's ability to recognize a small number of categories, which improves the classification accuracy in a wide range of visual tasks and helps the model to converge faster, especially when faced with the category imbalance problem. The loss function can be written as shown below:

$$FL(p_t) = -(1 - p_t)^{\gamma} log(p_t), \tag{5}$$

Here, a hyperparameter γ is added, which the authors call the focusing parameter (focusing parameter), and in the experiments in this paper, γ = 2 works best, while it is the standard cross-entropy function when it is 0.

4. Experiments

4.1. Datasets

We created a passport background texture printing image dataset for model training. The collection of images for the dataset consisted of two sources of acquisition: a library of passport samples from publicly accessible sources and self-collection. This dataset is one of the contributions of this work, as there is no publicly available dataset and we will describe its features in detail. The main passport sample repository from publicly accessible sources is the online Public Register of Authentic Travel and Identity Documents (PRADO). PRADO is an EU passport information repository that collects documents from the European Union and its member states as well as from countries that are part of the PRADO program. The online system is open to the public, and the sample repository collects a large number of images of passport security features, which provides a wealth of data for our study.

In the process of constructing the dataset, we not only utilized the rich resources provided by PRADO but also paid special attention to the various scenarios that may occur in practical applications, such as stamp occlusion, defacement, and changes in the presentation of the underlining texture under different light sources. In order to simulate these scenarios, we captured passport background texture images containing stamp blocking and defacement as well as photographed images of the backing texture under a variety of light source conditions to ensure that the dataset can cover more complex situations in practical applications. In addition, to increase the diversity of laser printing and inkjet printing underprints, laser and inkjet printers were used to generate additional laser printing and inkjet printing underprint images, which were then scanned into electronic data. These images not only enriched the variety of the dataset but also helped to improve the model's ability to recognize the underprints of these two printing methods.

All self-collected images of the dataset were collected using the VSC6000 document detector, and the inclusion criteria in data collection are as follows:

- Images must have a resolution exceeding 200 by 200 pixels.
- Photos should be taken using natural lighting.
- The object of the image should be Transformed design, lithographic printing, laser printing, or inkjet printing.

The images in the dataset can be categorized into four categories, namely "Transformed design", "Laser printing", "Lithographic printing", and "Inkjet printing". The labels of the four categories of images are set as 0, 1, 2, and 3. In our study, we first divided the dataset into a training set and a test set in the ratio of 8:2. Then, we used data augmentation (enhancement) to increase the diversity and richness of the dataset, thus improving the generalization ability of the model. As shown in Figure 4, we introduce a variety of image enhancement techniques, including random horizontal flip, vertical flip, random rotation, color dithering, random affine transformation, and perspective transformation. These operations can simulate the changes of images in different scenes, thus increasing the model's adaptability to various environments. As a result, the number of samples per category in the training set was increased to 1000, while in the test set, the number of samples per category was increased to 250.

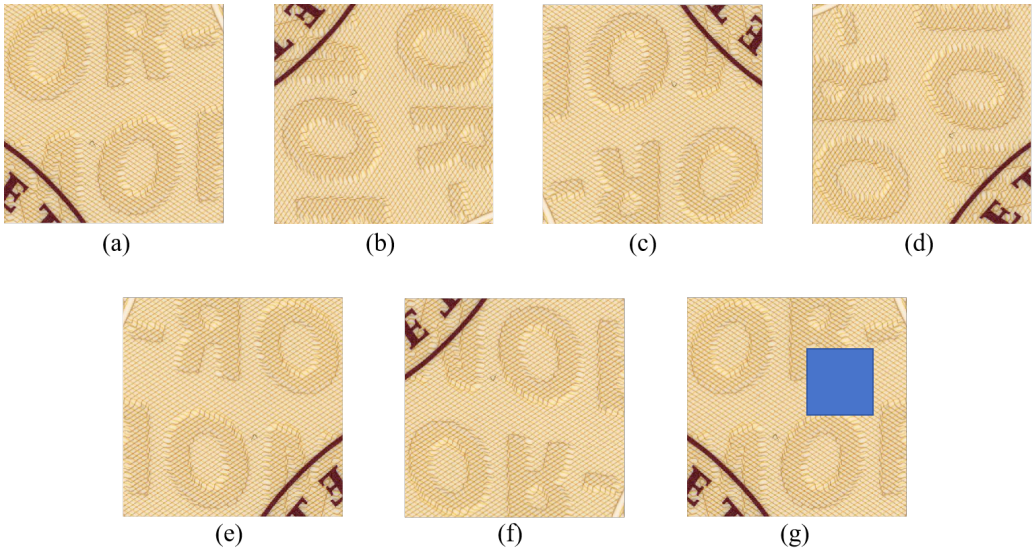


Figure 4. Seven samples of the dataset after data enhancement. (**a**) The original image, (**b**) the original image rotated 90°, (**c**) the original image rotated 180°, (**d**) the original image rotated 270°, (**e**) the original image flipped horizontally, (**f**) the original image flipped vertically, (**g**) and the original image with a random rectangle added.

4.2. Experiment Platform

The parameters for all models were optimized utilizing the Adam optimizer, beginning with a learning rate of 0.001 and employing a mini-batch size of 128. For the training images, we set the size of random clipping images to $224 \times 224 \times 3$. Training proceeded for 100 epochs for each model. The coding, which included data preprocessing and the implementation of the algorithms, was completed in Python. The models were run on an Ubuntu 18.04 LTS workstation, featuring an Intel i7-8700K CPU@3.70GHz and an NVIDIA GTX10800Ti GPU, using PyTorch version 1.7.

4.3. Evaluation Criteria

In this study, the assessment criteria encompass top-1 accuracy, precision, recall, F1 score, and network parameter count [44]. The corresponding formula is described as follows:

$$Accuracy = \frac{TN+TA}{TN+TA+FN+FA} \tag{6}$$

$$Precision = \frac{TA}{TA+FA} \tag{7}$$

$$Recall = \frac{TA}{TA+FN} \tag{8}$$

$$F_1 score = \frac{2 \times precision \times recall}{precision+recall} \tag{9}$$

in which *TN* stands for the number of nominal observations correctly predicted as nominal, *TA* stands for the number of abnormal observations correctly predicted as abnormal, *FN* stands for the number of abnormal observations wrongly predicted as nominal, and *FA* stands for the number of nominal observations wrongly predicted as abnormal.

Model classification accuracy is a pivotal metric for assessing the model's overall effectiveness.

Nevertheless, for this application, the model must operate within the constraints of a mobile device that has limited random access memory (RAM). We also consider the parameter count as an additional metric for evaluation.

4.4. Ablation Study

In order to evaluate the effectiveness and respective contributions of the modules in our proposed PBNet network, we conducted a series of stepwise ablation experiments on the dataset. The following models were compared to evaluate their performance:

- Model 1: MobileNet v2 is the experimental model.
- Model 2: Transformer is the experimental model.
- Model 3: A parallel structure is adopted, consisting of a CNN branch and a Transformer branch, where the CNN branch is the residual block of MobileNet v2 and the Transformer branch is the standard transformer encoder.
- Model 4: Based on model 3, we replaced the loss function with focal loss.
- Model 5: Based on model 4, we added the feature fusion module.

Table 1 provides a quantitative analysis of the experimental results of the different modules on the dataset. Our proposed model (Model 5) achieves AUC improvements of 5.8%, 16%, 2.7%, and 3.2% compared to Models 1 to 4. It can be seen that although MobileNet v2 is a lightweight network, it has good performance itself and still achieves an accuracy of 82.3%; however, Transformer can only achieve an accuracy of 72.1%.

Table 1. Results of ablation experiment.

Model	CNN	Transformer	Focal Loss	Feature Enhance Module	Accuracy	Precision	Recall	F1 Score
1	✓				0.8230	0.6671	0.6608	0.8201
2		✓			0.7210	0.7165	0.7160	0.7185
3	✓	✓			0.8540	0.8517	0.8500	0.8490
4	✓	✓	✓		0.8490	0.8450	0.8370	0.8380
5	✓	✓	✓	✓	0.8810	0.8829	0.8790	0.8799

As can be seen from the metrics in Table 1, accuracy rises from 85.4% in Model 3 to 88.1% in Model 5, and the F1 score rises from 0.8201 in Model 3 to 0.8799 in Model 5, which indicates that the use of these two improved methods superimposed on CNN-Transformer can obtain richer semantic feature information, and so Model 5 is chosen as the optimal model CNN-Transformer for this paper for experimentation.

The test results showed the following. (1) With an increase in the number of acceptable model parameters, the performance of the model combining the two improved strategies outperforms that of the model with one of the strategies alone. (2) Two improved strategies can be effectively integrated, and the improved model based on ShuffleNet v2 increases the accuracy of security feature recognition in complex context by 0.8 percentage points. (3) Although the accuracy of the model degraded after the introduction of focal loss alone (Models 3 and 4), we found that the combination of focal loss with the feature enhancement module was able to make up for this shortcoming (Models 4 and 5) and helped the model regain a higher level of accuracy. The combination of the two significantly improves the model's ability to discriminate difficult-to-classify samples, allowing it to exhibit stronger classification ability and better robustness when dealing with complex, unbalanced datasets. With this optimization strategy, our model achieves better performance in terms of precision,

recall, and F1 score, reflecting the important role of focal loss in improving the overall stability of the model and meeting the challenges of practical applications.

In this study, we employ the Grad-CAM model [45] to gain clearer insights into how the model allocates classification weights when it incorporates attention mechanisms for the task of classifying images of passport security features. The Grad-CAM technique reveals the regions of the image that capture the model's attention, leveraging gradient data from the final convolutional layer of the CNN to determine the significance of individual neurons in making targeted attention choices [46]. As an illustration, using models with various attention modules, Figure 5 displays the Grad-CAM visualizations for four distinct types of passport security feature images.

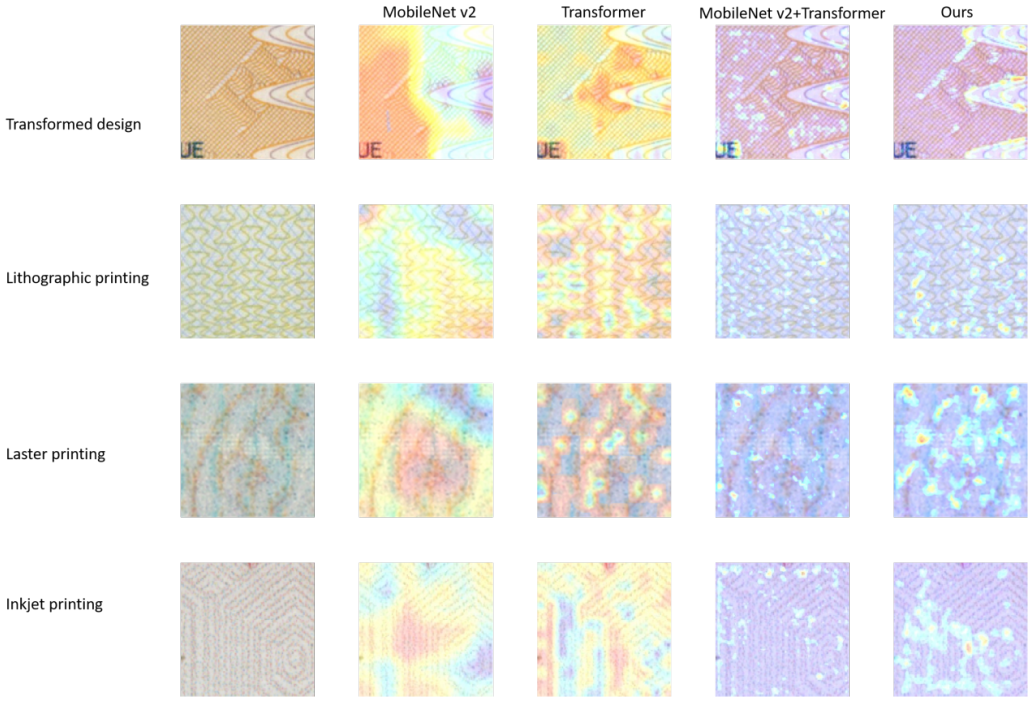


Figure 5. Grad-CAM maps of different models for four-classes passport background texture printing image classification.

4.5. Training Results of Model

Figure 6 shows the trend of loss and accuracy in 100 epochs of the training process. Whether it is the loss or the acc curve, the green curve demonstrates a steady trend, indicating that the model continues to optimize its parameters to better fit the data during the training process. Although there may be slight fluctuations or risk of overfitting at some stages, overall, it is a well-performing training process. Our model performance is positive.

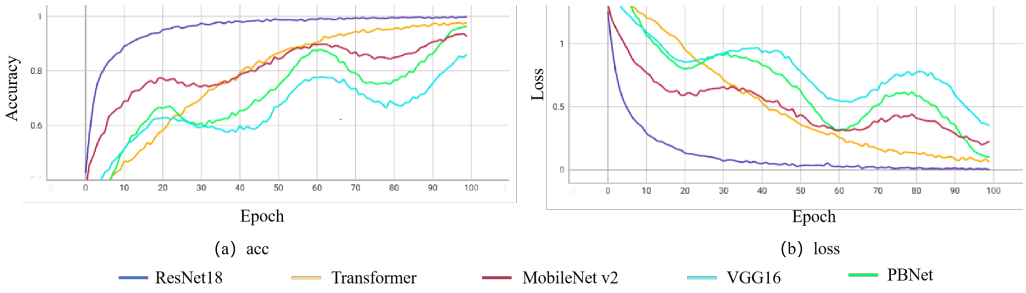


Figure 6. Comparison of training curves for 5 models.

In order to better evaluate the performance of the deep learning-based image classification model for passport background texture, this experiment introduces the confusion

matrix [47] to reveal the classification performance of the improved model. Figure 7 shows the recognition results of the passport background texture images test set before and after the improvement. This study sets labels corresponding to 0, 1, 2, and 3 for Transformed design, laser printing, lithographic printing, and inkjet printing. The horizontal coordinate indicates the labels of the predicted passport background texture types, and the vertical coordinate indicates the real passport background texture labels.

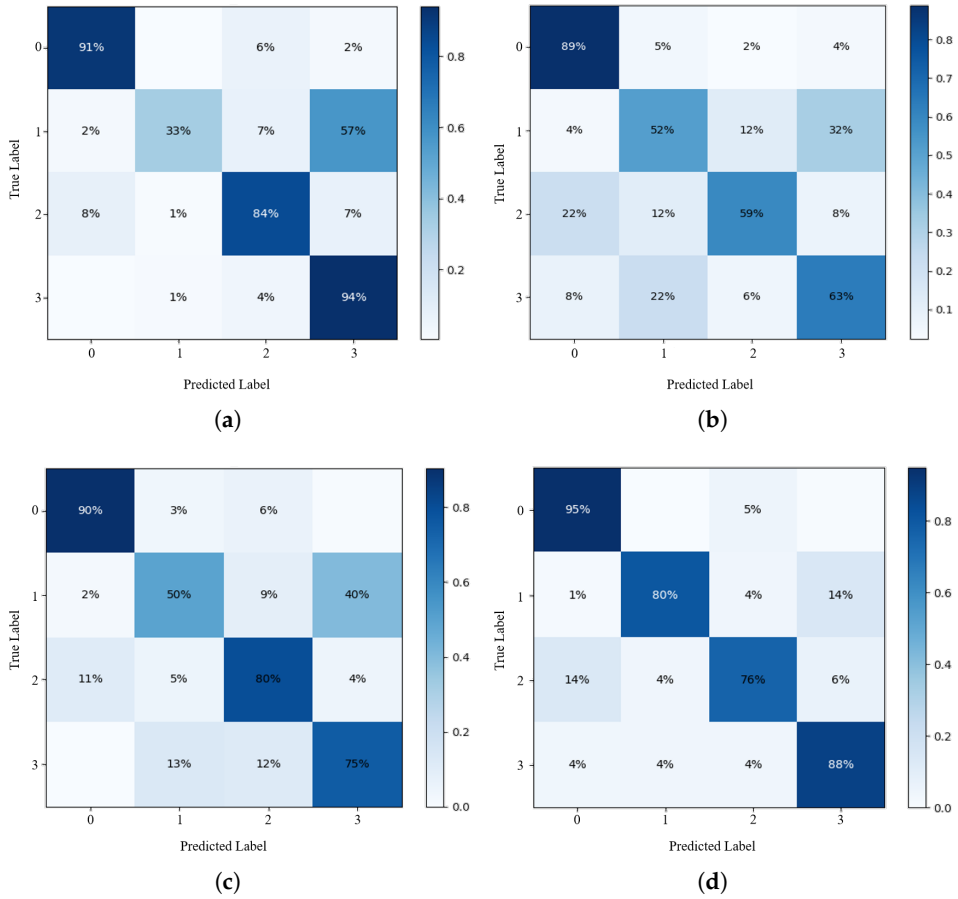


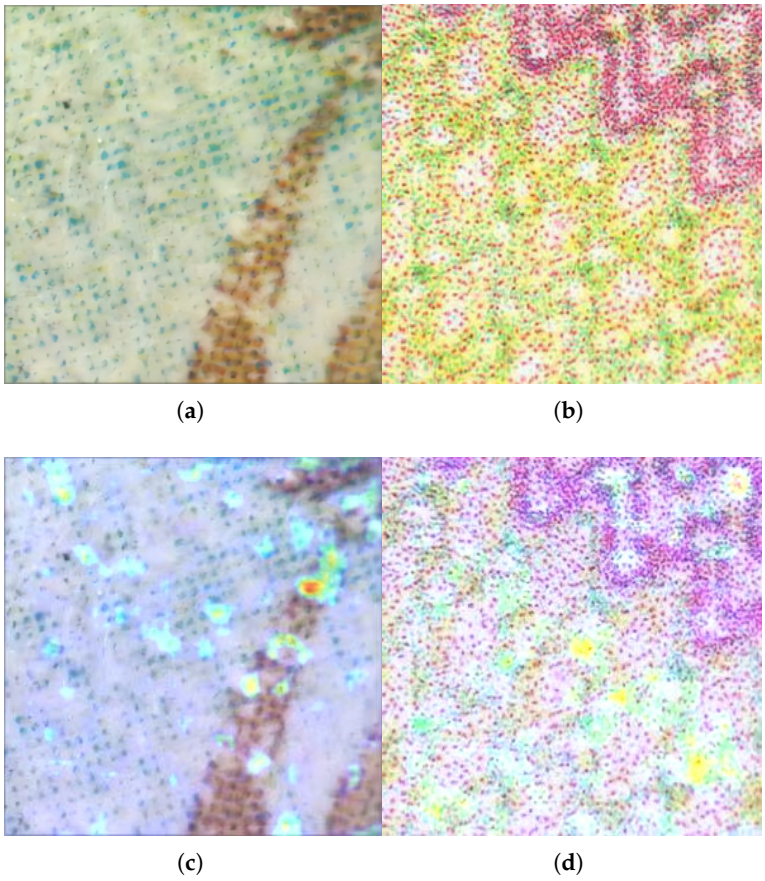
Figure 7. Confusion matrices of four models in the test set. (**a**) MobileNet v2. (**b**) Transformer. (**c**) MobileNet v2 + Transformer. (**d**) PBNet.

We can see that the number of images identified incorrectly by MobileNet v2, Transformer and MobileNet v2 + Transformer for label 1 is relatively high. The accuracy was only 33%, 52% and 50%; most of the laser printing images are misclassified as labeled 3 (inkjet printing). However, PBNet greatly improves the confusion between labels. This method increases the accuracy rate of label 3 to 80%, improving the accuracy of identifying the passport background texture test set effectively.

The feature enhancement module effectively improves the model's ability to recognize key features by combining channel and spatial attention mechanisms. Laser printing (label 1) and inkjet printing (label 3) were misclassified more often than other background textures. To explain the cause of this phenomenon, we extracted a portion of the misclassified images for observation and analysis.

Figure 8 shows two misclassified laser printing images. Both images use dots to form patterns, and they contain a large number of densely packed dots, creating a textured visual effect. The dots appear evenly distributed across the entire image with no obvious areas of concentration or sparsity, adding to their similarity. This makes it challenging to accurately categorize the images under particular lighting and background conditions. Moreover, from the two heatmaps, we can observe a high degree of similarity between the regions of interest for both laser printing and inkjet printing images, particularly in

terms of color distribution, the shape and location of the points of interest, which are concentrated in the highlighted areas and granular details. This similarity suggests that the model struggles to effectively differentiate between different printing methods when focusing on local details. When the two types of printed images exhibit similar patterns in minute texture and color features, the model tends to overlook their global characteristics, leading to incorrect classification. This indicates that the model may be over-relying on local similarities while failing to capture the essential differences during feature extraction.



(**a**) (**b**)

(**c**) (**d**)

Figure 8. Incorrectly classified samples in the passport background texture identification experiment. (**a**) Laser printing. (**b**) Inkjet printing. (**c**) Heatmap of laser printing. (**d**) Heatmap of inkjet printing.

For the purpose of visualizing the outcome of model recognition tests, a selection of passport background texture images was utilized. These images were fed into the operational PBNet. The classification results displayed in Figure 9 demonstrate that the model can still precisely determine the image categories regardless of the presence of intricate backgrounds or image blurriness. This highlights the model's robust generalization capability. Consequently, PBNet is confirmed as a proficient model for categorizing passport background textures, offering a commendable balance between recognition speed and accuracy.

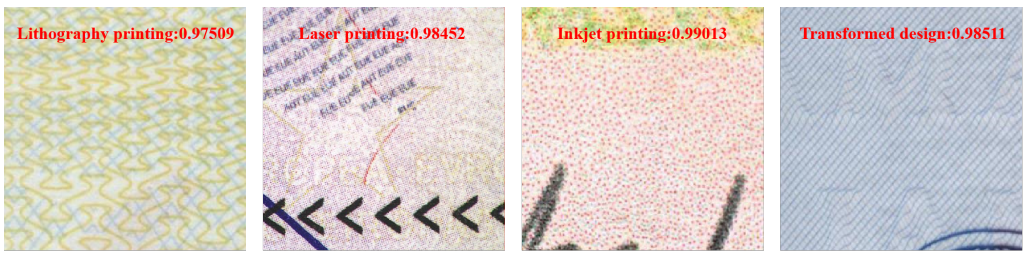


Figure 9. The visualization of model recognition results achieved satisfactory results.

4.6. Compare with State-of-the-Art Models

To demonstrate the effectiveness of the model introduced in this manuscript, we engaged in experimental comparisons with several prevalent models. We chose four classical deep learning models to compare our proposed model and other machine models: Vgg, Resnet18, Transformer, and MobileNet v2. These four models are common models in today's research. The comparison findings are shown in Table 2.

Table 2. Comparison of training results of each model on the self-built dataset.

Model	Train ACC (%)	Val ACC (%)	Number of Parameters	Inference Speed	FLOPs
VGG	86.75%	54.60%	138.35 M	0.0157	15.48 G
Resnet18	99.85%	69.00%	11.69 M	0.0111	1.82 G
MobileNet v2	92.83%	82.30%	3.50 M	0.0105	0.32 G
Transformer	96.33%	72.10%	85.80 M	0.0139	16.85 G
Swin-Transformer	98.18%	75.30%	27.52 M	0.0115	4.35 G
PBNet	94.67%	88.10%	1.05 M	0.0225	43.24 G

The experimental results show that although our model is slightly slower than MobileNet v2 and Resnet18 in terms of inference speed, which is 2.25 s, it is worth noting that in real-world application scenarios, the test time of 100 images taking 2.25 s is acceptable. On the other hand, the significant improvement of our model on Val ACC may make up for this lack of speed. Our model shows a significant advantage in Validation Set Accuracy (Val ACC) of 0.8810, which is a result that is the highest among all compared models. In terms of the number of parameters, our model is significantly lower than other models: VGG is as high as 138.35 M, Resnet18 is 11.69 M, the Transformer is 85.80 M, and even the lightweight architecture MobileNet v2 is 3.5 M. In comparison, our model is only 1.05 M, which is an advantage that makes PBNet excel in many key aspects such as storage efficiency, deployment convenience, computational efficiency, real-time processing capability, generalization capability, scalability, maintenance and update, and cost-effectiveness, which is especially suitable for resource-constrained environments and application scenarios that require fast response while reducing the maintenance and operational costs of the model and improving the overall performance and practicality.

On the other hand, we use the FLOPs to compare the computational complexity of different deep learning models. It can be seen from Table 2 that even though the FLOPs of our model (PBNet) are larger than all the other models, it did not prevent the model from achieving more than 85% of accuracy. Through the analysis of the experimental results above, we have demonstrated that lightweight models excel in application scenarios requiring rapid response, achieving an ideal balance between accuracy and inference time. This finding aligns with the research of Lee et al. [48] in Watt-effnet, which emphasizes the efficiency of lightweight models in image classification tasks. Similarly, Kyrkou and Theocharides [49] in Emergencynet proposed an efficient model for aerial image classification that also confirms the practicality of lightweight architectures in emergency monitoring. These results collectively support our viewpoint that lightweight models are an optimal choice for achieving an ideal balance between precision and efficiency.

5. Conclusions

Since Transformer has made great strides in NLP, it has been used in a wide variety of tasks. In recent years, researchers have attempted to utilize Transformer-based methods to solve visual tasks due to their strong long-range dependency representation capabilities. Convolutional Neural Networks (CNNs), which can effectively extract contextual information and spatial details, have been widely used in visual tasks. However, it is more difficult for neural networks to extract long-range dependencies than Transformers, which are not good at extracting spatially relevant information and maintaining spatial details. To an extent, these two properties of CNNs and Transformers are complementary.

To efficiently compensate for the shortcomings of CNNs and Transformers, we propose a CNN and Transformer Complementary Network (PBNet) for passport background texture image classification. Our PBNet has two encoders: a CNN and a Transformer. The CNN encoder is constructed by MobileNet v2 to extract features in the CNN domain, while the Transformer one is created to capture long-range dependent features. The feature enhancement module uses the channel enhancement and spatial enhancement channel to combine these features from CNN and Transformer domains effectively. The Transformer decoder is created by Swin Transformer blocks for further improving long-range representation and recovering feature maps to the input size. In our experiments, we found that PBNet struck a good balance between accuracy and inference speed. First, PBNet achieved an accuracy of 88.1% on the validation set, which was significantly better than ResNet18 (69.0%), MobileNet v2 (82.3%), and swin-transformer (75.3%), indicating that it performed well in terms of accuracy. Second, the number of parameters of PBNet was only 1.05 M, which was the lightest among all comparative models, and it significantly reduced the model complexity compared to large-scale models such as VGG (138.35 M) and Transformer (85.80 M). Furthermore, although the inference time of PBNet was 0.0225 s, which was slightly higher than that of MobileNet v2 (0.0105 s), this inference time was still in the acceptable range in the context of higher accuracy, and compared to the inference times of VGG (0.0157 s) and Transformer (0.0139 s), PBNet provided higher accuracy while maintaining a reasonable inference speed. Even with the slight increase in inference time, PBNet's performance is still within acceptable limits in real-world usage scenarios. In a typical boundary control or document validation process, processing a few seconds per batch of images (e.g., 100 images) is a reasonable trade-off to ensure accurate recognition. In real-world applications, the balance between accuracy and inference time depends on the specific requirements of the task. For mobile deployments in passport verification, accuracy is often prioritized, as errors in identifying forged documents can have serious security implications. Therefore, our PBNet model aims to optimize both accuracy and efficiency without making significant sacrifices in either direction.

However, despite our considerable efforts in dataset construction and model training, there are still some limitations. (1) The diversity of the current dataset is still insufficient. Although we have tried our best to collect passport background texture images from multiple sources and employed a variety of data enhancement techniques, the various complexities that may arise in real-world applications, such as the diversity and complexity of forged passports and the variation of the passport background texture under different lighting and breakage conditions, are still not adequately covered. This may lead to limitations in the recognition accuracy and generalization ability of the model when facing certain specific scenarios. (2) Although our model has achieved certain results in preliminary experiments, there is still room for further improvement in its accuracy. With the continuous progress of technology and the deep development of applications, the requirements on model performance will become higher and higher. Therefore, we need to keep exploring new techniques and methods to improve the accuracy and robustness of the model to better meet the demands of practical applications.

To address these limitations, we plan to take the following steps to improve our future work. Firstly, we will continue the diversity and complexity of texture collection. We plan to collect passport background texture images from more diverse sources, especially including samples of forged passports, to further enrich the diversity and complexity of the dataset. At the same time, we will also include more scenarios that may occur in real-world applications, such as side-lighting conditions and broken condition underlining, in order to more fully simulate real-world environments. In addition, we will explore techniques such as Generative Adversarial Networks (GANs) that use generative models to synthetically generate fake passport background textures to further increase the diversity and complexity of the dataset. Secondly, we will try some other methods such as class re-sampling or other loss functions (e.g., weighted cross-entropy) to improve the performance of the model on minority classes. Thirdly, we will try to conduct experiments using more advanced

models. With the continuous development of deep learning technology, various new model architectures and algorithms are emerging [50]. We will pay close attention to the development of these new technologies and try to apply them to our research. Through continuous experimentation and optimization, we expect to further improve the accuracy and robustness of the models to better meet the needs of practical applications.

Author Contributions: Conceptualization, J.X. and T.Z.; Data curation, J.X. and D.J.; Formal analysis, J.X., T.Z. and J.W.; Funding acquisition, J.X. and T.Z.; Investigation, J.X.; Methodology, Z.L. and T.Z.; Project administration, J.X. and Z.L.; Supervision, T.Z. and J.W.; Validation, D.J., J.X. and L.T.; Visualization, J.X. and L.T.; Writing—review and editing, J.X.; Writing—original draft, J.X. and T.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science Research Project of Hebei Education Department (No. QN2022199).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Baechler, S. Document Fraud: Will Your Identity Be Secure in the Twenty-First Century? *Eur. J. Crim. Policy Res.* **2020**, *26*, 379–398. [CrossRef]
2. Devlin, C.; Chadwick, S.; Moret, S.; Baechler, S.; Raymond, J.; Morelato, M. The Potential of Using the Forensic Profiles of Australian Fraudulent Identity Documents to Assist Intelligence-Led Policing. *Aust. J. Forensic Sci.* **2023**, *55*, 720–730. [CrossRef]
3. Moulin, S.L.; Ertan, E.; Martin, D.; Baechler, S. Cross-Border Forensic Profiling of Fraudulent Identity and Travel Documents: A Pilot Project Between France and Switzerland. *Sci. Justice* **2024**, *64*, 202–209. [CrossRef] [PubMed]
4. Moulin, S.L.; Weyermann, C.; Baechler, S. An Efficient Method to Detect Series of Fraudulent Identity Documents Based on Digitised Forensic Data. *Sci. Justice* **2022**, *62*, 610–620. [CrossRef] [PubMed]
5. Leese, M.; Noori, S.; Scheel, S. Data Matters: The Politics and Practices of Digital Border and Migration Management. *Geopolitics* **2022**, *27*, 5–25. [CrossRef]
6. Saadi, Z.M.; Sadiq, A.T.; Akif, O.Z.; Farhan, A.K. A Survey: Security Vulnerabilities and Protective Strategies for Graphical Passwords. *Electronics* **2024**, *13*, 3042. [CrossRef]
7. Ouassam, E.; Dabachine, Y.; Hmina, N.; Bouikhalene, B. Improving the Efficiency and Security of Passport Control Processes at Airports by Using the R-cnn Object Detection Model. *Baghdad Sci. J.* **2024**, *21*, 0524–0536. [CrossRef]
8. Elebe, T.M.; Kurnaz, S. Efficient Detection of Refugees and Migrants in Turkey Using Convolutional Neural Network. *Phys. Commun.* **2023**, *59*, 102078. [CrossRef]
9. Liu, Y.; Joren, H.; Gupta, O.; Raviv, D. Mrz Code Extraction from Visa and Passport Documents Using Convolutional Neural Networks. *Int. J. Doc. Anal. Recognit. (IJDAR)* **2022**, *25*, 29–39. [CrossRef]
10. Dimitriou, E.; Michailidis, N. Printable Conductive Inks Used for the Fabrication of Electronics: An Overview. *Nanotechnology* **2021**, *32*, 502009. [CrossRef]
11. Tao, Y.M.; Tang, H.; Yang, X.; Chen, X.H. Assessment of High-Quality Counterfeit Stamp Impressions Generated by Inkjet Printers via Texture Analysis and Likelihood Ratio. *Forensic Sci. Int.* **2023**, *344*, 111573. [CrossRef] [PubMed]
12. Li, T.; Zhang, Z.; Zhu, M.; Cui, Z.; Wei, D. Combining Transformer Global and Local Feature Extraction for Object Detection. *Complex Intell. Syst.* **2024**, *10*, 4897–4920. [CrossRef]
13. Zhao, Z.; Chen, T.; Dou, J.; Liu, G.; Plaza, A. Landslide Susceptibility Mapping Considering Landslide Local-Global Features Based on Cnn and Transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 7475–7489. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
15. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
17. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
18. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. Squeezenet: Alexnet-Level Accuracy with 50x Fewer Parameters and <0.5 Mb Model Size. *arXiv* **2016**, arXiv:1602.07360.
19. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

20. Burri, S.R.; Ahuja, S.; Kumar, A.; Baliyan, A. Exploring the Effectiveness of Optimized Convolutional Neural Network in Transfer Learning for Image Classification: A Practical Approach. In Proceedings of the 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 5–6 May 2023; pp. 598–602.
21. Han, Q.; Qian, X.; Xu, H.; Wu, K.; Meng, L.; Qiu, Z.; Weng, T.; Zhou, B.; Gao, X. Dm-cnn: Dynamic Multi-Scale Convolutional Neural Network with Uncertainty Quantification for Medical Image Classification. *Comput. Biol. Med.* **2024**, *168*, 107758. [CrossRef]
22. Zhang, J.; Wang, W.; Che, Q. Innovative Research on Intelligent Recognition of Winter Jujube Defects by Applying Convolutional Neural Networks. *Electronics* **2024**, *13*, 2941. [CrossRef]
23. Jiang, P.; Xue, Y.; Neri, F. Convolutional Neural Network Pruning Based on Multi-Objective Feature Map Selection for Image Classification. *Appl. Soft Comput.* **2023**, *139*, 110229. [CrossRef]
24. Zhang, M.; Liu, L.; Jin, Y.; Lei, Z.; Wang, Z.; Jiao, L. Tree-Shaped Multiobjective Evolutionary Cnn for Hyperspectral Image Classification. *Appl. Soft Comput.* **2024**, *152*, 111176. [CrossRef]
25. Pandiri, D.K.; Murugan, R.; Goel, T. Smart Soil Image Classification System Using Lightweight Convolutional Neural Network. *Expert Syst. Appl.* **2024**, *238*, 122185. [CrossRef]
26. Li, X.; Liu, Y.; Zheng, L.; Zhang, W. A Lightweight Convolutional Spiking Neural Network for Fires Detection Based on Acoustics. *Electronics* **2024**, *13*, 2948. [CrossRef]
27. Ran, Q.; Zhou, Y.; Hong, D.; Bi, M.; Ni, L.; Li, X.; Ahmad, M. Deep Transformer and Few-Shot Learning for Hyperspectral Image Classification. *CAAI Trans. Intell. Technol.* **2023**, *8*, 1323–1336. [CrossRef]
28. Zhou, W.; Dou, P.; Su, T.; Hu, H.; Zheng, Z. Feature Learning Network with Transformer for Multi-Label Image Classification. *Pattern Recognit.* **2023**, *136*, 109203. [CrossRef]
29. Wu, L.; Zhou, J.; Jiang, H.; Yang, X.; Zhan, Y.; Zhang, Y. Predicting the Characteristics of High-Speed Serial Links Based on a Deep Neural Network (DNN)—Transformer Cascaded Model. *Electronics* **2024**, *13*, 3064. [CrossRef]
30. Wang, D.; Zhang, J.; Du, B.; Zhang, L.; Tao, D. Dcn-t: Dual Context Network with Transformer for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2023**, *32*, 2536–2551. [CrossRef] [PubMed]
31. Gong, L.Y.; Li, X.J.; Chong, P.H.J. Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector. *Electronics* **2024**, *13*, 3045. [CrossRef]
32. Yuan, F.; Zhang, Z.; Fang, Z. An Effective Cnn and Transformer Complementary Network for Medical Image Segmentation. *Pattern Recognit.* **2023**, *136*, 109228. [CrossRef]
33. Zhang, X.; Su, Y.; Gao, L.; Bruzzone, L.; Gu, X.; Tian, Q. A Lightweight Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–17. [CrossRef]
34. Al-Ghadi, M.; Ming, Z.; Gomez-Krämer, P.; Burie, J.C. Identity Documents Authentication Based on Forgery Detection of Guilloche Pattern. *arXiv* **2022**, arXiv:2206.10989.
35. Sirajudeen, M.; Anitha, R. Forgery Document Detection in Information Management System Using Cognitive Techniques. *J. Intell. Fuzzy Syst.* **2020**, *39*, 8057–8068. [CrossRef]
36. Gonzalez, S.; Valenzuela, A.; Tapia, J. Hybrid Two-Stage Architecture for Tampering Detection of Chipless ID Cards. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *3*, 89–100. [CrossRef]
37. Ghanmi, N.; Nabli, C.; Awal, A.M. Checksim: A Reference-Based Identity Document Verification by Image Similarity Measure. In *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, 5–10 September 2021, Proceedings, Part I 16*; Springer: Cham, Switzerland, 2021; pp. 422–436.
38. Xu, J.; Jia, D.; Lin, Z.; Zhou, T. Psfnet: A Deep Learning Network for Fake Passport Detection. *IEEE Access* **2022**, *10*, 123337–123348. [CrossRef]
39. Jeny, A.A.; Junayed, M.S.; Atik, S.T. Passnet-Country Identification by Classifying Passport Cover Using Deep Convolutional Neural Networks. In Proceedings of the 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 21–23 December 2018; pp. 1–6.
40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
41. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
42. Dina, A.S.; Siddique, A.; Manivannan, D. A Deep Learning Approach for Intrusion Detection in Internet of Things Using Focal Loss Function. *Internet Things* **2023**, *22*, 100699. [CrossRef]
43. Batool, U.; Shapiai, M.I.; Mostafa, S.A.; Ibrahim, M.Z. An Attention-Augmented Convolutional Neural Network with Focal Loss for Mixed-Type Wafer Defect Classification. *IEEE Access* **2023**, *11*, 108891–108905. [CrossRef]
44. Bono, F.M.; Radicioni, L.; Cinquemani, S. A novel approach for quality control of automated production lines working under highly inconsistent conditions. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106149. [CrossRef]
45. Selvaraju, R.R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; Batra, D. Grad-cam: Why Did You Say That? *arXiv* **2016**, arXiv:1611.07450.
46. Liu, Y.; Zhang, Z.; Liu, X.; Lei, W.; Xia, X. Deep Learning Based Mineral Image Classification Combined with Visual Attention Mechanism. *IEEE Access* **2021**, *9*, 98091–98109. [CrossRef]

47. Valero-Carreras, D.; Alcaraz, J.; Landete, M. Comparing Two SVM Models Through Different Metrics Based on the Confusion Matrix. *Comput. Oper. Res.* **2023**, *152*, 106131. [CrossRef]
48. Lee, G.Y.; Dam, T.; Ferdaus, M.M.; Poenar, D.P.; Duong, V.N. Watt-effnet: A lightweight and accurate model for classifying aerial disaster images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
49. Kyrkou, C.; Theocharides, T. EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 1687–1699. [CrossRef]
50. Wang, F.; Liang, Y.; Lin, Z.; Zhou, J.; Zhou, T. SSA-ELM: A Hybrid Learning Model for Short-Term Traffic Flow Forecasting. *Mathematics* **2024**, *12*, 1895. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.