QS STARS RATED FOR EXCELLENCE 2012

FPT Education

FPT UNIVERSITY

Department of Computer Science

FPT University

Extraction Information from Vietnamese ID Card Images

Do Cong Duy	HE150348
Vu Doan Quang	HE153583
Vu Hoang Tai Toan	HE150224

Supervised by

A thesis submitted in part fulfillment of the degree of BSc. in Computer Science with the supervision of M.S. Do Thai Giang

December, 2023

Student's Declaration

We declare that this project titled "*Extraction Information from Vietnamese ID Card Images*", submitted as requirement for the award of degree of Bachelors in Computer Science, does not contain any material previously submitted for a degree in any university; and that to the best of our knowledge, it does not contain any materials previously published or written by another person except where due reference is made in the text.

We understand that the management of Department of Computer Science, FPT University, has a zero tolerance policy towards plagiarism. Therefore, We, as authors of the above-mentioned thesis, solemnly declare that no portion of our thesis has been plagiarized and any material used in the thesis from other sources is properly referenced.

We further understand that if we are found guilty of any form of plagiarism in the thesis work even after graduation, the University reserves the right to revoke our BS degree.

Do Cong Duy Signature: ________________

Vu Doan Quang Signature: ________________

Vu Hoang Tai Toan Signature: ________________

____

Verified by Plagiarism Cell Officer
Dated:

Acknowledgements

We extend our heartfelt appreciation to M.S. Do Thai Giang for his invaluable guidance and support throughout this research endeavor. His advice from the initial stages of model development to providing relevant resources has been instrumental in shaping the direction of our study. We also extend our appreciation to the faculty members of the AI Department at FPT University for their expertise and dedication, which have created an enriching research environment. Their valuable insights and discussions have greatly contributed to the development of our research. We are extremely grateful for the support and resources of FPT University, which have been essential for the completion of this research project. Additionally, we express our gratitude to our colleagues and fellow researchers within the AI Department for their collaboration and assistance during this endeavor. After all, we would like to send our gratitude to our family and friends. They have always encouraged and been a great motivation to help us complete our university journey.

Do Cong Duy
Vu Doan Quang
Vu Hoang Tai Toan

Abstract

The efficient extraction of information from ID cards is vital for various daily services, such as legal, banking, insurance, and medical processes. Nevertheless, in numerous developing countries nations like Vietnam, this task is predominantly manual, resulting in time-consuming, monotonous, and error-prone processes. This thesis presents a deep learning system specifically designed to extract information from images of Vietnamese ID cards. The proposed system involves three sequential steps: ID card alignment algorithm, text detection and text recognition. The initial step incorporates two neural networks, YOLACT and ResNet50 for segmentation and classification model, alongside an image processing technique. The second step employs YOLOv7 for text detection. The third step is to utilize VietOCR based on Attention OCR for recognizing Vietnamese optical text on the cards. In experimental evaluations, the proposed system demonstrates a notable reduction in processing time, especially in scenarios where corners of the ID card are obstructed. It effectively addresses challenges when compared to existing methodologies, achieving a high mAP@[0.5:0.95] for box is 97.21 % and mask is 99.58 % for ID card segmentation, 74.0 % for text detection. The system also exhibits exceptional precision score for full-sentence with recognition rates of 98.19 % for Vietnamese optical texts and 97.60 % for the ID card classification model. Overall, our implementation of the proposed method achieves an average system-wide accuracy rate about 97.98 %.

Keywords: OCR, ID Card, Vietnamese ID Card, Information Extraction.

Contents

1. INTRODUCTION **5**

1.1 Background 5

1.2 Literature Review 6

1.3 Motivation 6

1.4 Objectives and Contribution 7

1.5 Organization 7

2. RELATED WORK **8**

2.1 Document localization approaches 8

2.2 Text detection approaches 9

2.3 Text recognition approaches 9

3. PROPOSAL METHODOLOGY **11**

3.1 ID Card alignment algorithm 11

3.1.1 Segmentation model 11

3.1.2 Alignment algorithm 14

3.1.3 Classification model 16

3.2 Text detection 18

3.2.1 Overview YOLOv7 18

3.2.2 YOLOv7's Architecture 18

3.3 Text recognition 22

4. IMPLEMENT AND ANALYSIS **24**

4.1 Dataset 24

4.2 Results 26

5. CONCLUSION 29

References 35

List of Figures

1 System pipeline. 11

2 YOLACT Architecture [1]. 12

3 Prototype mask. 13

4 Head Architecture. 13

5 Segmentation mask image applied Contour and Convex hull. 15

6 Segmentation mask image cluster points are obtained. 16

7 Segment mask image with 4 points found. 17

8 Alignment ID card image. 17

9 YOLOv7's Backbone. 20

10 YOLOv7 compound scaling [2]. 20

11 Implement to ELAN block. 21

12 Planned re-parameterized model. 21

13 YOLOR Implicit Knowledge. 22

14 TranformerOCR with image. 23

15 Visual data annotation. 25

16 Visual data labels. 25

17 The number of losses on training dataset. With B: box loss, C: class loss, M: mask loss, S: segmentation loss and T: total loss. 26

18 Mean Average Precision (mAP) of segmentation model (YOLACT model). 26

19 Accuracy and Loss on Classification model (Transfer learning ResNet50). 28

20 Web-app interface. 28

List of Tables

1 Evaluating the text detection model (YOLOv7 model). 27

2 Precision full-sentence comparison between pre-trained and transfer learning of VietOCR model. 27

1. INTRODUCTION

1.1 Background

In the vibrant tapestry of Vietnam's progress, digital transformation weaves the threads of innovation, connecting a nation's past to its future. Embracing technology, Vietnam dances towards a future where connectivity transforms challenges into opportunities, and where the digital rhythm propels the nation into a new era of growth and prosperity.

As one of the fastest-growing economies in Southeast Asia, Vietnam has embraced the power of digital technologies to propel itself into the forefront of innovation and progress. This transformation is not merely about adopting new tools but signifies a fundamental shift in mindset, a recognition that the fusion of technology and traditional values can unlock unprecedented opportunities for growth and development.

In the dynamic realm of information technology, identity cards (ID Cards) have transcended their role as mere proof and important of identification for Vietnamese citizens. They have evolved into cruScial components, seamlessly integrated into various facets of daily life, particularly in the telecommunications, healthcare, and hospitality sectors. The emergence of electronic Know Your Customer (eKYC) systems within the banking sector has further amplified the importance of automatic information extraction from identity cards, establishing it as a standardized protocol for individual verification [3]. Despite this growing recognition, the prevailing method of manually extracting information from such cards remains a time-consuming, laborious process, susceptible to errors. Due to the requirements of developing new technologies, researchers have started to learn and develop algorithms that can extract data and they have found promising results.

1.2 Literature Review

Identification cards (ID cards) are ubiquitous in modern society, serving as a means of verifying an individual's identity. These cards typically contain various personal information, including name, date of birth, address, and photograph. Extracting this information from ID card images is becoming increasingly important, and there are many different applications with many different models and architectures to extract information from ID card images. Extracting information from ID card images is divided into three phases including ID card localization, text detection, and text recognition (OCR). One of the most common approaches in early phases is to use object detection to identify the 4 corners of the ID card, then based on the coordinates of the corners on the original image to crop the ID card and apply text detection and text recognition (OCR) to extract information as output [4] [5] [6] the above research results achieved an average higher than 91.0 %. In addition, there are many traditional methods in image processing that can be applied in the early phases to crop the ID card as a rectangle for examples: geometrical transformations and the Hough Transform in [7] [8]. Another sample techniques in image processing was Martin Hirzer proposed [9] then next step still uses text detection and text recognition (OCR), accuracy of the whole system is 94.0 % and 95.28 %.

1.3 Motivation

Although it is clear that the above studies have made great contributions to the field but there are still some certain limitations. Firstly, in case more than one corner is missing in the above studies and not being able to identify the 4 corners of the ID card, it is impossible to cut the ID card based on the coordinates of the corners on the original photo. Secondly, in reality not all areas are rectangular. In fact, when taking a citizen identification card from a phone from different angles, most of the results will be trapezoids and parallelograms, so it cannot be done according to traditional methods in computer vision.

1.4 Objectives and Contribution

Our research contribution lies in the second aspect. Firstly, because there is no public data set on images of Vietnamese ID Cards, the contribution of our collected dataset to other researchers can be used. Second, our proposed methods will improve other research limitations. That can work well in cases where many angles are missing and the ID card photo is in fact not rectangular due to the angle when the photo was taken.

1.5 Organization

The remaining sections of the paper are structured as follows: Section 2 offers an overview and reviews previous research on information extraction from images. Section 3 delves into the architecture details and alignment algorithm. Section 4 covers the experimental evaluation. Lastly, Section 5 concludes with a discussion of the system.

2. RELATED WORK

Information extraction or document extraction is the task of automatically extracting information from images, videos, sounds and documents that is of interest as a problem to be solved. Historically, information extraction systems were first published in the mid-1980s [10]. After nearly half a century of technological development, especially the development of machine learning, deep learning and today's information or document extraction methods also known as image to text extraction or optical character recognition (OCR) can be divided into several steps depending on the complexity of the task and the desired accuracy and case to apply in real life. General breakdown of the process: document localization, text detection and text recognition.

2.1 Document localization approaches

In the document localization that correspond to documents such as passports, invoices, ID cards or receipts. There are several common solutions are employed for document localization from images including image processing techniques and machine learning, deep learning methods. With image processing techniques, edge detection [11] techniques identify discontinuities in the intensity or color of an image using algorithms like Sobel, Canny or Laplacian [12] and Connected component analysis [13] [14] identifies connected regions in an image, typically using algorithms like flood-filling or union-find. These connected regions can represent documents in the image. In tradition methods [15] [16] key points are information extraction and characterized by a description method like Scale-Invariant Feature Transform for the sample to match the feature. Although, above methods handle success on plastic cards it is not good feature match case occurs when handling paper cards with complex patterns and color transitions. Machine learning, deep learning techniques have been extensively applied to solve the problem of document localization from images, offering promising results in terms of accuracy and robustness. According to reference [17], the corners of the document are first predicted using Convolutional Neural Networks. This method can works correctly on this task but also raises the limit for some cases in where partial or entire corners are covered.

2.2 Text detection approaches

Text detection in ID cards images the regular methods based on image processing techniques is blurred to eliminate noise, then changed to binary form images to completely removed the background of image [18] [19] and filter out non-informative text using morphological methods [20] [21]. However, it's so hard to set a threshold suitable for varying degrees of lightness. Furthermore, color items in old ID cards can fade because of above reasons the effective approach for better accuracy is text detection as objection detection [22] based on RetinaNet model [23]. Text detection like object detection algorithms have been developed. There are two phase detectors and one phase detector with their own advantages and disadvantages. For example, with two stage detectors, it involves generating regions of interest in the first stage and classifying these regions in the second stage [24] [25]. With a one-stage detector that directly predicts object classes from the image without creating region of interest, this proposed method provides to achieve faster run-time [26].

2.3 Text recognition approaches

Text recognition also known as optical character recognition (OCR) is process of converting text embedded within images into machine-readable text format. Text recognition in ID card images is a crucial task for various applications including identity verification, document automation and border control. The complex nature of ID card images with varying layouts, fonts and potential distortions, poses challenges for text recognition algorithms. Several research papers have addressed these challenges. Based on image processing techniques, each character copped into a single image, recognized by template matching [19]. Character segmentation in text processing leads to bad results making it complicated. To improve performance, several research papers have proposing novel techniques and advancements in text recognition for ID card images uses deep learning methods particularly Convolutional Neural Networks have emerged as powerful methods for text recognition in ID card image data [27] [28], enabling accurate text recognition even challenging scenarios but this time-consuming method. Attention mechanisms have been incorporated into deep learning models to improve text recognition in ID card images [29] [30]. These mechanisms allow the model to focus on relevant regions of the image, such as the text areas, and suppress noise or irrelevant background information. Reference [29] reported significant improvements in recognition accuracy compared to traditional text recognition methods. The proposed attention-based model achieved an average recognition accuracy of 96.1 % outperforming with other methods by margin of up to 7 %.

3. PROPOSAL METHODOLOGY

The proposed approach consists of three stages which are ID card alignment algorithm, text detection and text recognition as Figure 1 depicts the pipeline of the proposed system and the output of each process. It encompasses a range of conventional image processing, machine learning, and deep learning techniques employed in the field of computer vision.

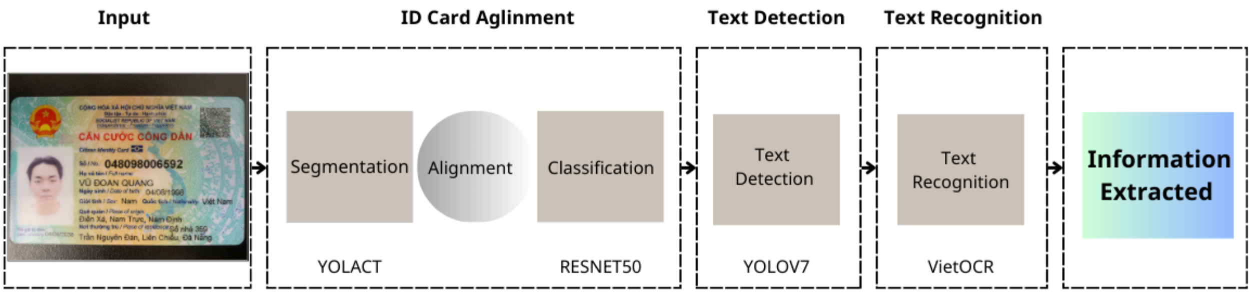


Figure 1: System pipeline.

3.1 ID Card alignment algorithm

Because of the unrestricted nature of the captured images, three pre-processing steps were implemented to prepare them for subsequent stages. These steps include segmentation, alignment, and classification. These tasks enable the separation of cards from the background image and the vertical alignment of the cards.

3.1.1 Segmentation model

Instance segmentation and semantic segmentation are both computer vision techniques that aim to identify and localize objects in images. However, there are some key differences between the two techniques that make instance segmentation more suitable for certain applications. Semantic segmentation [31] [1] is task of identifying and locating all pixels in an image that belong to a particular object category this means that semantic segmentation can only distinguish between different object categories. Instance segmentation [32] on the other hand is the task of identifying and locating individual instance of objects in an image even if they are overlapping or partially occluded and instance segmentation provides more detailed information about objects in an image as it produces object-level segmentation masks that delineate the boundaries of each object instance. Above are the reasons deciding to use instance segmentation model. However, traditional instance segmentation methods are often slow and computationally expensive. In this study, present YOLACT (You Only Look At Coefficient Ts) model [33] a real-time instance segmentation model that achieves state-of-the-art performance while maintaining high processing speed. YOLACT is a one-stage-detector and the architecture is shown in Figure 2 [1].

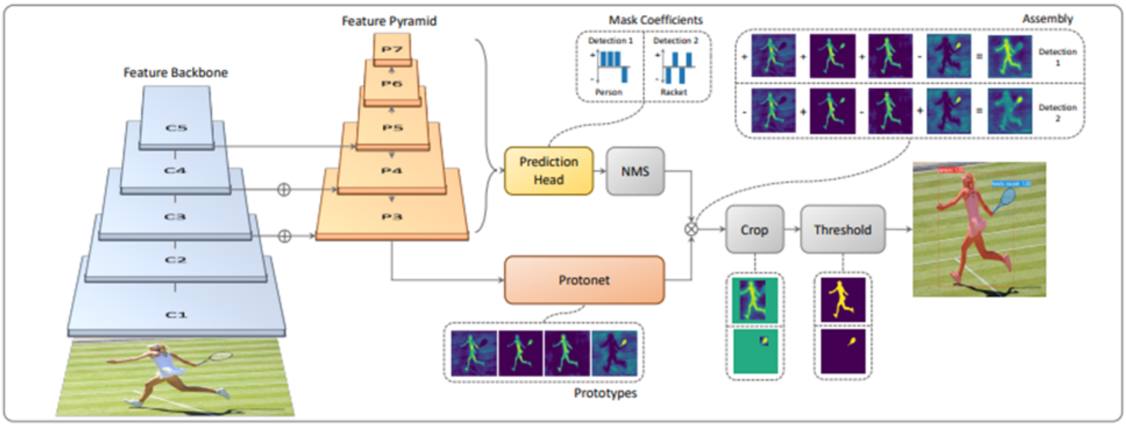


Figure 2: YOLACT Architecture [1].

YOLACT have been based on architecture of RetinaNet [34]. The instance segment work was then divided into two separate, simple, parallel branches: the prototype net branch using FCN [35] and the prediction head branch. Splitting into 2 branches like this helps optimize and parallelize calculations, helping this model achieve real-time speed, 3 - 5 times faster than current models.

Prototype Generation

The prototype generation branch (prototype net) predicts a set of K prototype masks for the entire image as shown in Figure 3. Implement prototype net as FCN. To create a more robust masks, a higher resolution prototypes would carry again both higher quality masks and better performance on smaller objects used P3 in FPN [36] because its largest feature layers are the deepest. The number K does not depend on the number of classes, but is optimized and selected after many trials. One point to note is that the larger K does not mean the better the output quality, because only the first prototype mask number affects the mask of objects, the remaining prototype mask numbers do not have much impact, just noise.

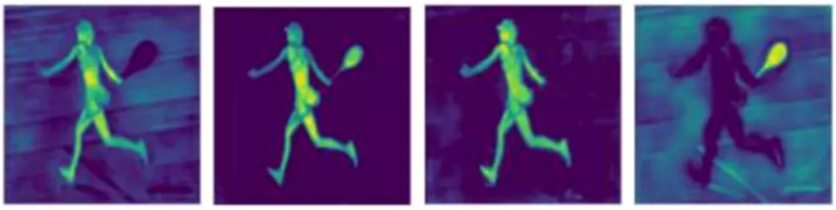


Figure 3: Prototype mask.

Prediction head

Prediction head has two branches in this process to generate the mask coefficients: one branch to predict C class confidences and other branch to predict 4 bounding box regression. To predict mask coefficient, we simply add a third branch in parallel to predicts k mask coefficients as shown in Figure 4, one corresponding to each prototype. Therefore, instead of generating 4 + C coefficients per each anchor YOLACT generate 4 + C + K. The Tanh function is applied to the prediction k mask coefficients to produce a more stable outputs under non-linearity conditions.

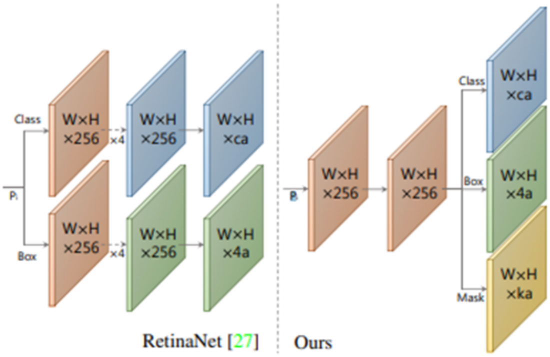


Figure 4: Head Architecture.

Mask Assembly

To generate instance masks combine work of the Prototype net and Prediction head is to combine the prototype masks and mask coefficients with the use of linear combination 1. These operations can be performed efficiently using a single matrix multiplication with sigmoid:

$$M = \sigma(PC^T) \tag{1}$$

where P is an $h \times w \times k$ matrix of prototype mask and C is the $n \times k$ matrix of mask coefficients.

Losses use three losses to train the YOLACT model for instance segmentation: classification loss L_{cls}, box regression loss L_{box}, and mask loss L_{mask}. L_{cls} and L_{box} are similar to those in [31], and L_{mask} 2 is the pixel-wise binary cross entropy that measures the difference between the predicted segmentation mask M and the true segmentation mask M_{gt}:

$$L_{\text{mask}} = \text{BCE}(M, M_{\text{gt}}) \tag{2}$$

3.1.2 Alignment algorithm

The output of segmentation model is a instance segmentation mask image. It is typically represented as a binary image where each pixel is either 0 (background) or 1 (objects). These alignment steps are important for processing the raw output of first model causing the segmentation mask image represent a pixel region of any object shape. Would like to present an alignment algorithm that have experimented. The goal of this algorithm is to find a quadrilateral with smallest area containing the area of ID card. OpenCV [1] already provides the minRectArea function, however not all areas are rectangles in facts most are trapezoids and parallelograms. To find a quadrilateral we must find its 4 vertices. First, I use Contour to find a path where all the points at the boundary of the object create a concave polygon surrounding the object. Due to the presence of obscured corners in actual ID card images, the Contour algorithm will inevitably produce a substantially larger set of boundary points. Decided to use Convex hull in OpenCV to reduce the points at the boundary and create a line will form a convex polygon surrounding the object as Figure 5.

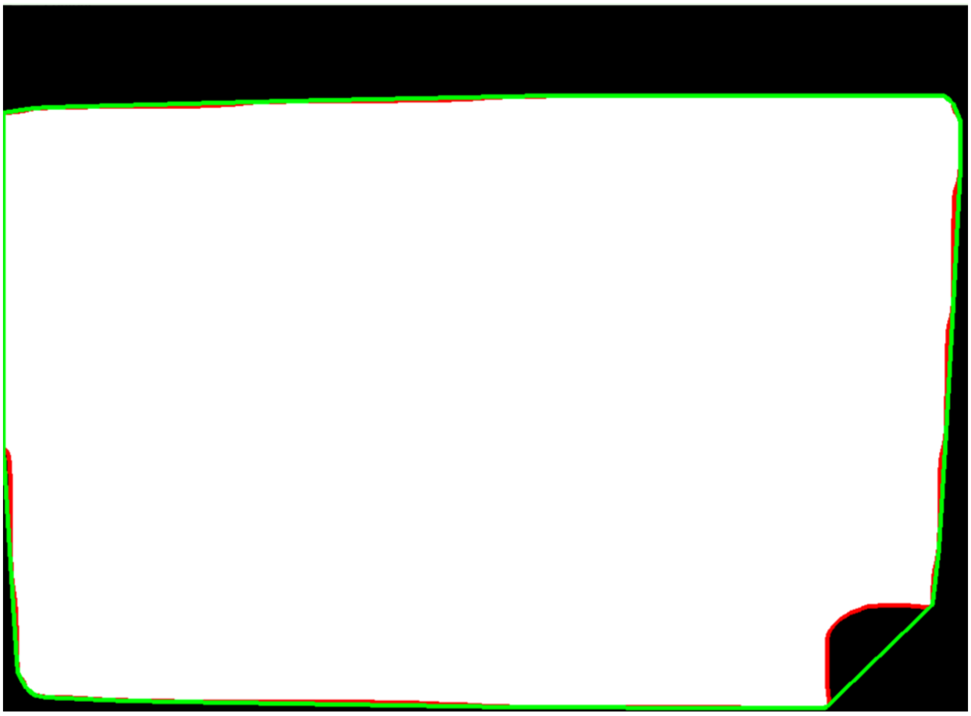


Figure 5: Segmentation mask image applied Contour and Convex hull.

It is necessary to remove points that are almost collinear and keep the points located on corners by calculating the angle at each point with two points close to it according to the formula 3 below with a threshold set at 170 degree.

$$\cos\theta = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|} \tag{3}$$

Finally, I obtained clusters of points located at the corners of the ID card as shown in Figure 6. Based on these points, you will find the 4 largest edges created by the previously found points. These 4 edges will be the 4 edges of the citizen identification card. Then the intersection of the 4 longest sides just found are the 4 vertices of the quadrilateral to be found.

After finding 4 points of the quadrilateral containing the ID card as shown in Figure 7. Apply the getPerspectiveTransform in OpenCV function to find a linear transformation matrix between 2 quadrilaterals (mask image and original image) and the target quadrilateral is a rectangle with size 1026x640 (the correct size for the following OCR step) as shown in Figure 8. Therefore, after testing we noticed the image had distorted or pen-

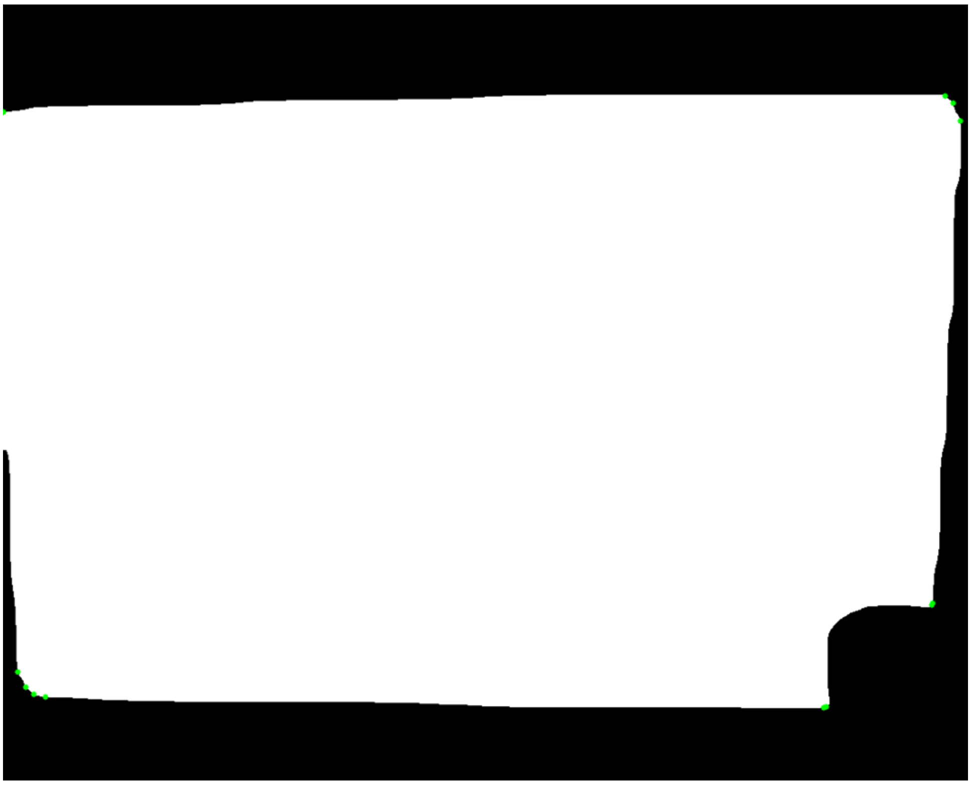
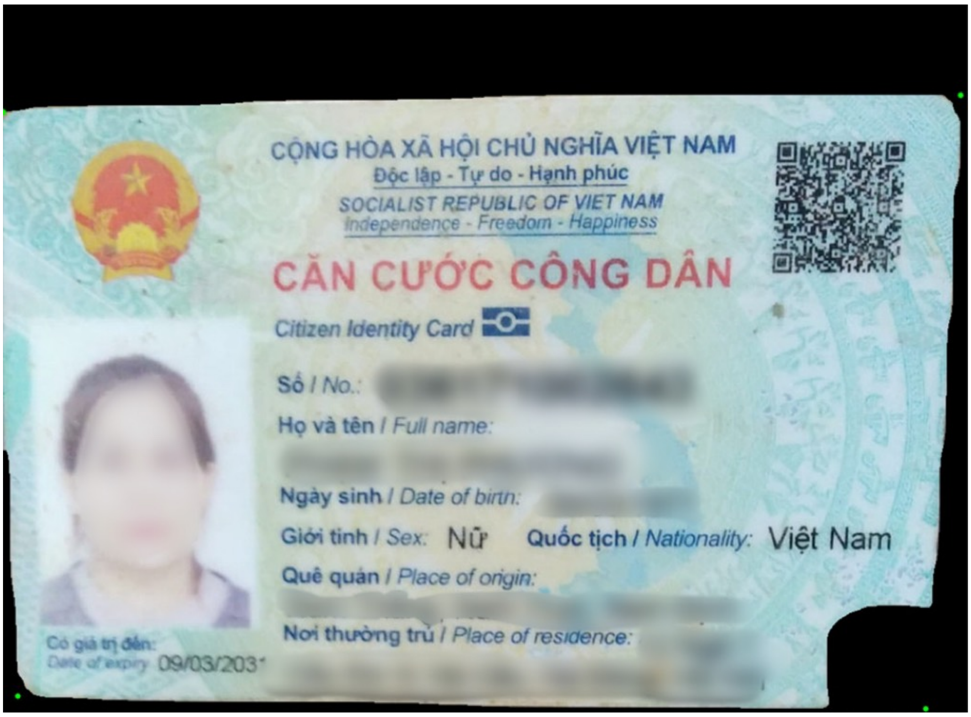


Figure 6: Segmentation mask image cluster points are obtained.

tagonal shape. So this approach Alignment algorithm solution still works correctly. If the image is distorted enough to produce 4 inaccuracies then the image does not provide the information needed for the next OCR step (e.g., covering the hand up to two-thirds of the card).

3.1.3 Classification model

Thanks to transfer learning, in the image classification task decided to approach transfer learning pre-trained ResNet50 model [32] for specific tasks leveraging their pre-learned knowledge and reducing training time while improving performance. ResNet50 is a 50-layer CNN that has revolutionized the field of image classification. Its innovative architecture addresses the vanishing gradient it to efficiently handle complex image features and achieve high performance on computer vision task. Understanding the architecture number of layers with a depth of 50 layers can learn intricate relationship and hierarchical representations within images leading to superior feature extraction with residual connec-



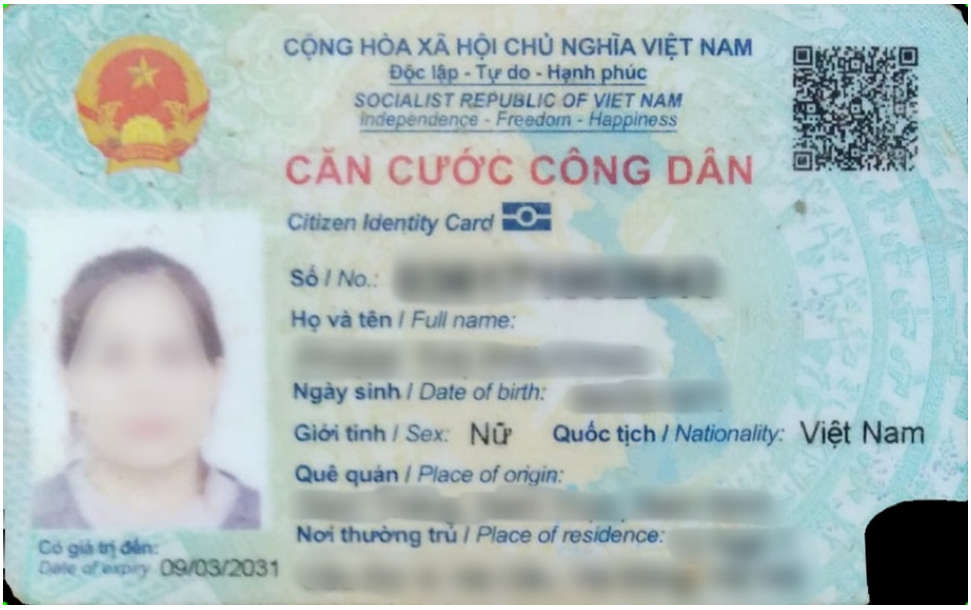
CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
SOCIALIST REPUBLIC OF VIET NAM
Independence - Freedom - Happiness

CĂN CƯỚC CÔNG DÂN
Citizen Identity Card

Số / No.: [redacted]
Họ và tên / Full name:
[redacted]
Ngày sinh / Date of birth: [redacted]
Giới tính / Sex: Nữ Quốc tịch / Nationality: Việt Nam
Quê quán / Place of origin:
[redacted]
Nơi thường trú / Place of residence: [redacted]
[redacted]

Có giá trị đến:
Date of expiry 09/03/203

Figure 7: Segment mask image with 4 points found.



CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
SOCIALIST REPUBLIC OF VIET NAM
Independence - Freedom - Happiness

CĂN CƯỚC CÔNG DÂN
Citizen Identity Card

Số / No.: [redacted]
Họ và tên / Full name:
[redacted]
Ngày sinh / Date of birth: [redacted]
Giới tính / Sex: Nữ Quốc tịch / Nationality: Việt Nam
Quê quán / Place of origin:
[redacted]
Nơi thường trú / Place of residence: [redacted]
[redacted]

Có giá trị đến:
Date of expiry 09/03/2031

Figure 8: Alignment ID card image.

tion ResNet50 incorporates shortcut connections that bypass some layers this mitigates the vanishing gradient problem.

3.2 Text detection

3.2.1 Overview YOLOv7

You Only Look Once (YOLO) stands out as a prominent object detection algorithm in computer vision, pioneering a streamlined approach to real-time object detection. Developed by Joseph Redmon and Santosh Divvala in 2016 [37], YOLO operates on the principle of conducting object detection in a single pass through the neural network, distinguishing itself from traditional two-stage detectors.

The central mechanism entails dividing the input image into a grid, with each grid cell predicting multiple bounding boxes along with a confidence score indicating the probability of object presence. Importantly, YOLO also predicts class probabilities for the identified objects within each bounding box. This unified prediction process eliminates the need for multiple passes and region proposals, contributing to YOLO's reputation for speed in real-time applications. YOLO has undergone various iterations, each refining the architecture for improved performance. Anchor boxes were introduced in YOLOv2 to enhance bounding box predictions, and subsequent versions, such as YOLOv3, further augmented the model's depth and accuracy.

Balancing speed and accuracy, YOLO finds applications in diverse fields, including autonomous vehicles, security surveillance, and image and video analysis. Its open-source implementations in frameworks like Darknet, TensorFlow, and PyTorch empower researchers and developers to leverage its capabilities, adapting it to meet specific requirements and trade-offs between speed and precision. YOLO's continuous evolution underscores its significance in advancing real-time object detection in the realm of computer vision.

3.2.2 YOLOv7's Architecture

YOLOv7 [2] inherits and follows the same architecture as its predecessors, using a convolutional neural network (CNN) to extract features from input images and predict bounding boxes and class probabilities for objects are detected. The network consists of several layers of backbone, neck, and probe.

Backbone

First, the input image undergoes processing through the Stem block, a block that transforms the image into feature maps consisting of two layers: a 3x3 convolutional layer and max pooling. Then, the feature maps continue into the ELAN block. ELAN [38] is a core architecture composed of three main components: Cross Stage Partial (CSP), Computation Block, and PointWiseConv [39]. The CSP component originates from YOLOv4 [40] with additional branches, and the computation block contains conv layers that perform computations to generate new features through a 3x3 convolution. Subsequently, the generated feature maps are merged at the end using the concatenation operator along the channel axis. This concatenated result then undergoes a PointWiseConv (1x1 conv) operation. The transition block is situated between the ELAN blocks. Each transition block reduces the size of the feature maps by a factor of 2 and consists of two layers: a 1x1 convolutional layer and max pooling. The overall YOLOv7 Backbone is shown in Figure 9.

Compound Model Scaling

In the case of the Computation Block inside the ELAN Block, the author of YOLOv7 noticed that when scaling the depth (increasing the number of layers in the computation block), the number of channels generated during concatenation also increases. Therefore, the author proposed a new scaling method called compound model scaling, which is illustrated in the Figure 10.

After some calculations, the author of YOLOv7 determined that by scaling the depth by 1.5 times, the width will be scaled by 1.25 times as well. Applying this depth and width scaling method to the ELAN Block, the number of layers in the computation block of the ELAN Block increases from 4 to 6 (1.5 times), and the number of channels during concatenation will increase from 64x4 to 64x5 (1.25 times), as shown in the Figure 11.

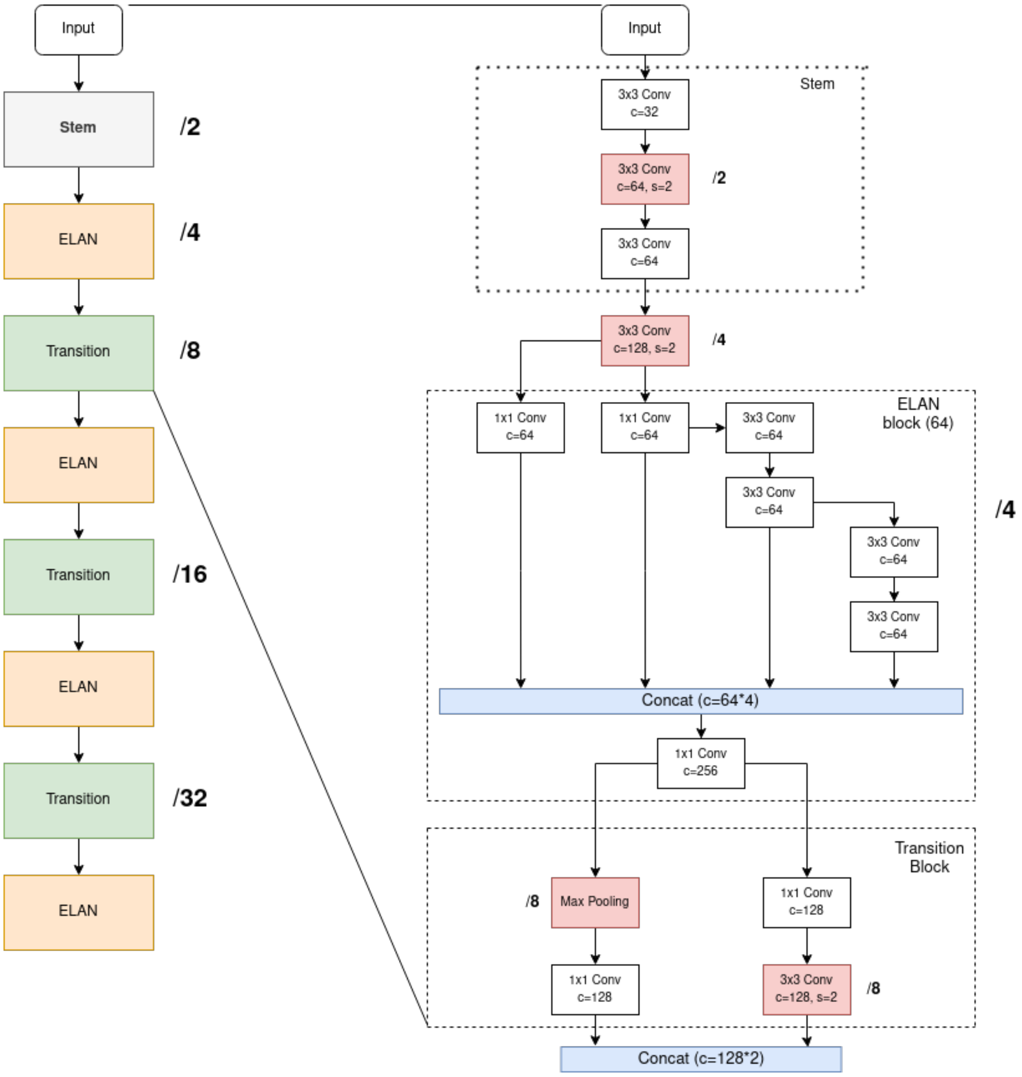


Figure 9: YOLOv7's Backbone.

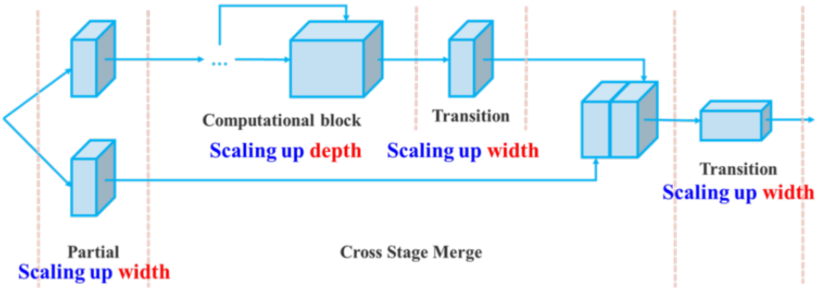


Figure 10: YOLOv7 compound scaling [2].

Re-parameterization Convolution (RepConv)

The feature maps from different scales are further processed with 3x3 convolutions corresponding to each scale. In YOLOv7, RepConv is applied based on RepVGG [41], which

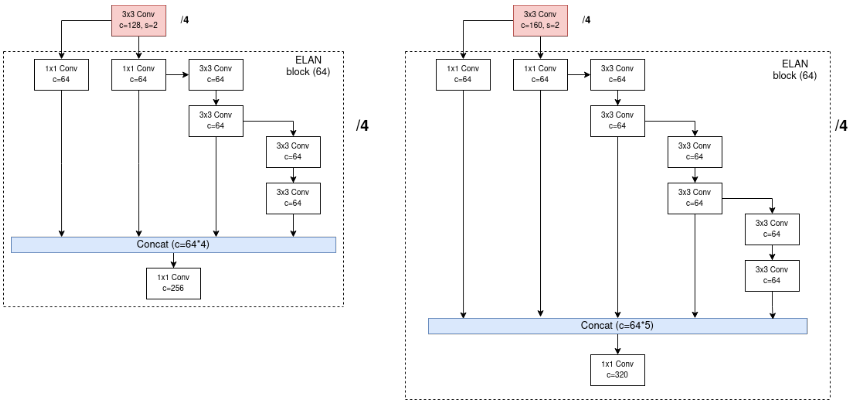
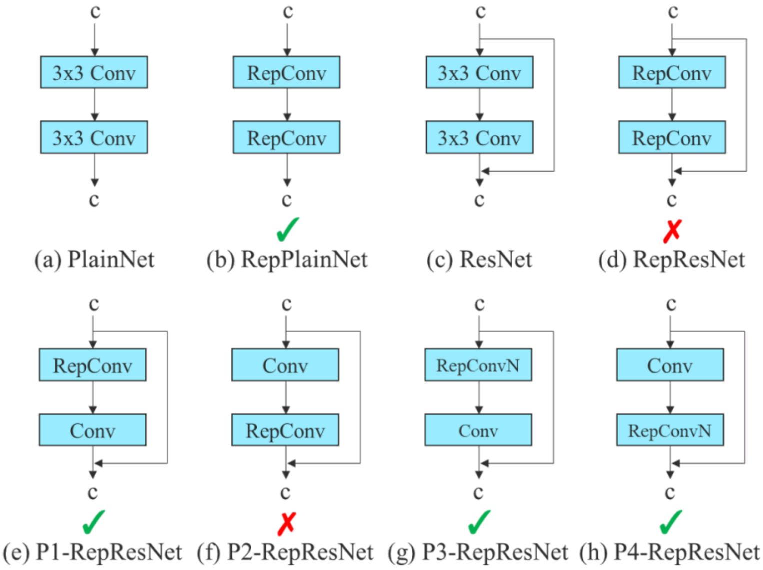


Figure 11: Implement to ELAN block.

consists of convolutional layers with a 3x3 kernel size combined with each other, along with a 1x1 identity connection. Through thorough analysis of RepConv and its performance compared to other architectures, the authors found that the identity connections can diminish the distinctive features of the feature maps. Therefore, they proposed RepConvN, which removes the identity connections, as illustrated in the Figure 12.



(a) PlainNet (b) RepPlainNet (c) ResNet (d) RepResNet

(e) P1-RepResNet (f) P2-RepResNet (g) P3-RepResNet (h) P4-RepResNet

Figure 12: Planned re-parameterized model.

Implicit Knowledge

At the output of the feature map, the author applies Implicit Knowledge from YOLOR [42], as depicted in the Figure 13. Here is an explanation of the concepts:

Explicit Knowledge: It refers to the knowledge that the model learns through exposure to specific labeled input images containing objects. The YOLOv7 network architecture extracts important features from the images, which are then used for object classification and localization.

Implicit Knowledge: It refers to the knowledge that the model deduces during the training process to discover the distinctive features of objects. There are three ways to represent this knowledge: vector, neural network, and combination. These representations are combined using two operations: addition and multiplication.

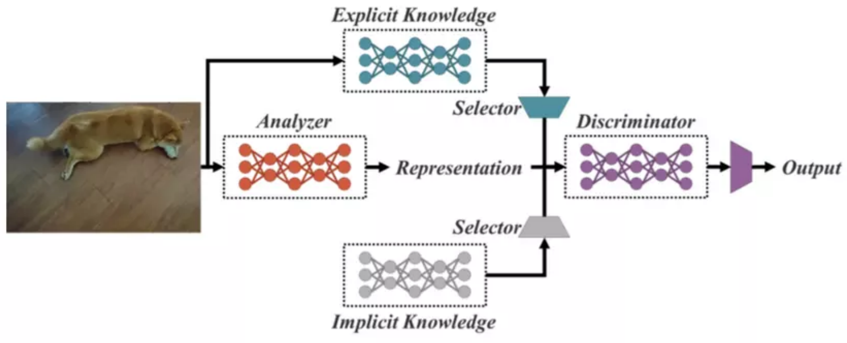


Figure 13: YOLOR Implicit Knowledge.

3.3 Text recognition

VietOCR

VietOCR library was built by author Pham Ba Cuong Quoc based on the Transformer OCR model. Unlike other language models, VietOCR's input will be an image instead of text. An image passed through the CNN model will produce a feature map with dimensions width x height x channel (W x H x C) show in Figure 14 [6]. The Transformer architecture relies on the attention mechanism, which enables the model to concentrate on significant portions of the input data for task execution. By employing a sequence of self-attention layers, the Transformer model processes the input information and produces contextual representations for individual words in the sentence. A distinguishing characteristic of the Transformer is its utilization of multi-head attention layers. These layers empower the model to learn how to attend to distinct aspects of the input, thereby enabling it to capture intricate relationships among words in a sentence.

The model was trained by the author on a dataset of over 10 million images, resulting in a Precision full sequence of 88%[1]. However, the model exhibits sensitivity to minor variations in the input when the dataset is not sufficiently trained. And, the use of beam search during sentence generation did not significantly improve the accuracy. Unlike text translation, where the information is explicitly provided, image-based information entails more uncertainty and requires additional inference.

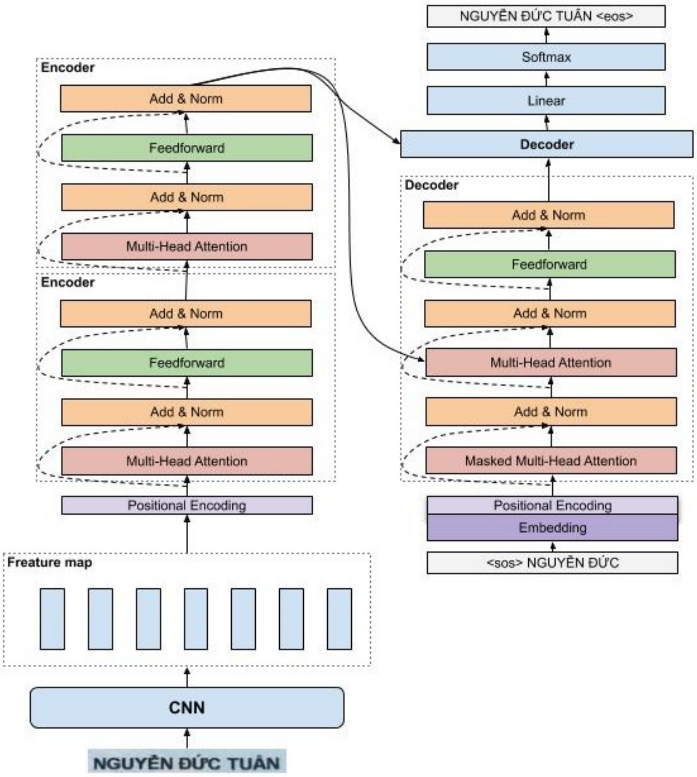


Figure 14: TranformerOCR with image.

[1] pbcquoc.github.io/vietocr

4. IMPLEMENT AND ANALYSIS

4.1 Dataset

Because ID cards are sensitive and personal information, not public data as mentioned before. Collecting a large data set is also difficult so our group is collected on social networks and especially from our friends and relatives. The study utilized four distinct datasets for model training and evaluation:

ID Cards Dataset: The dataset consists of 644 Vietnamese ID card images. To address privacy concerns, data augmentation techniques were utilized to augment the dataset. These techniques involved randomly rotating the original images.

Synthetic ID Card Images: To enhance the training dataset for the classifier model, half of the ID cards from the ID Cards dataset were selected and rotated by 180 degrees. This resulted in an additional half of the original dataset size.

Synthetic Text Images: To train the text area detection model, 500 alignment images were extracted from the output of the alignment card processing step. These alignment images were generated using the output model from the previous stage.

Synthetic Cropped Images: To transfer learning the OCR model, 8000 cropped images were generated from the 500 alignment images obtained in the previous step. These cropped images were assigned to 16 different classes.

Implementation detail

Dataset to train segmentation model (YOLACT model) using the form of the COCO dataset [43]. There are many tools support to annotation dataset with format of the COCO dataset. In this proposal method decided to used labelme is one of the most popular tools supporting annotation dataset in machine learning, deep learning. However, to minimize the time spent typing labels, can take the max min coordinates of the segment area to draw the bounding box. The labelme2coco.py [43] file does this automatically as shown in Figure 15.

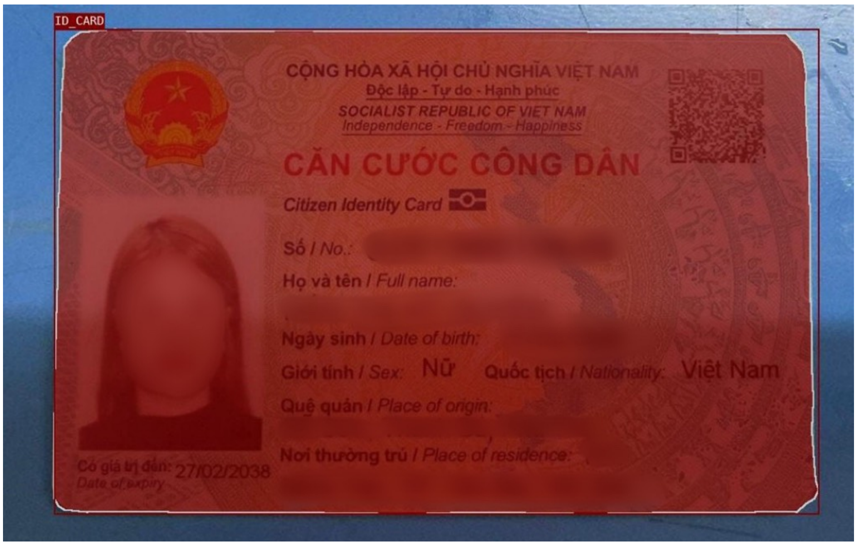

ID_CARD
CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
SOCIALIST REPUBLIC OF VIET NAM
Independence - Freedom - Happiness
CĂN CƯỚC CÔNG DÂN
Citizen Identity Card
Số / No.:
Họ và tên / Full name:
Ngày sinh / Date of birth:
Giới tính / Sex: Nữ
Quốc tịch / Nationality: Việt Nam
Quê quán / Place of origin:
Nơi thường trú / Place of residence:
Có giá trị đến: 27/02/2038
Date of expiry

Figure 15: Visual data annotation.

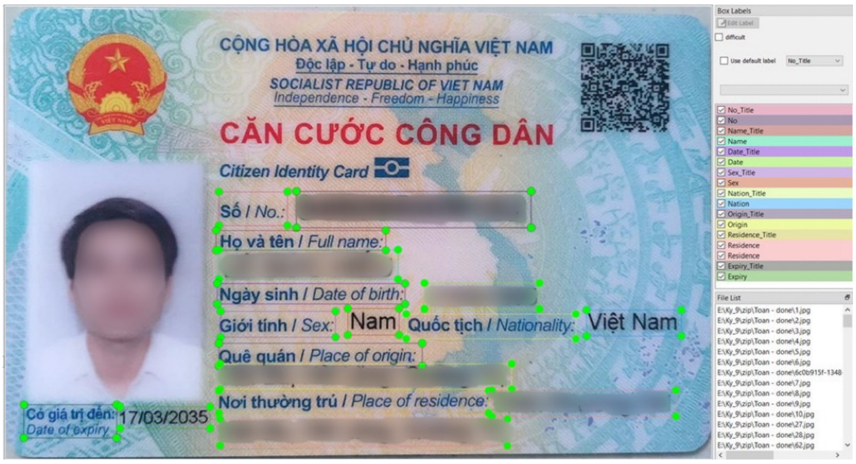


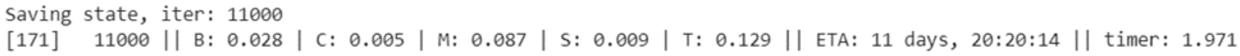
Figure 16: Visual data labels.

With the dataset for the text detection model (YOLOv7 model) used labelImg for the label bounding box for the 16 text classes on the front-facing ID card image as shown in Figure 16. Finally, prepare the data set for the OCR transfer learning model in the form "path_to_file_name[tab]labels" then save to the txt file.

4.2 Results

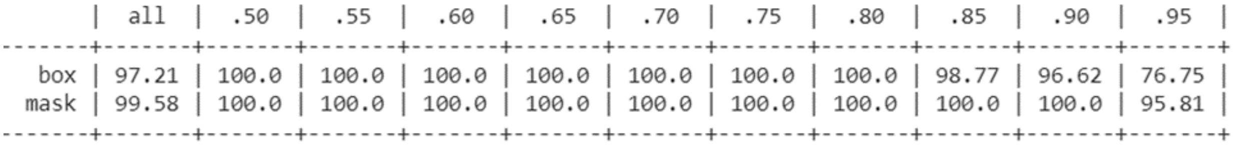
All models are implemented using Python 3.9 on NVIDIA A100 Tensor Core GPU. As mentioned previously, without publicly available data on ID card images, comparing accuracy with other studies is unfair. So in this study only focus on our data. Specifically, the training of this segmentation model stopped at 171 epochs with 11000 inters with the number of losses on the training data set shown in Figure 17. And the model also achieving mAP@[0.5:0.05:0.95] for box is 97.21 % and mask is 99.58 % as shown in Figure 18

.



```
Saving state, iter: 11000
[171]   11000 || B: 0.028 | C: 0.005 | M: 0.087 | S: 0.009 | T: 0.129 || ETA: 11 days, 20:20:14 || timer: 1.971
```

Figure 17: The number of losses on training dataset. With B: box loss, C: class loss, M: mask loss, S: segmentation loss and T: total loss. .



```
       |  all  |  .50  |  .55  |  .60  |  .65  |  .70  |  .75  |  .80  |  .85  |  .90  |  .95  |
-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
   box | 97.21 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 98.77 | 96.62 | 76.75 |
  mask | 99.58 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 95.81 |
-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+
```

Figure 18: Mean Average Precision (mAP) of segmentation model (YOLACT model).

Another model text detection model (YOLOv7 model) achieves mAP@[0.5:0.05:0.95] to 74 % for all classes as shown in Table.1. The effectiveness of the VietOCR transfer learning model is evaluated based on comparison results with the pre-trained model from scratch. Achieve an precision score of full-sentence for all classes around 98.19 % improvement with the pre-trained model is only about 67.15 % as Table.2 describes.

Table 1: Evaluating the text detection model (YOLOv7 model).

Class	**P**	**R**	**mAP@.5**	**mAP@.5:.95**
All	98.5%	96.80%	96.10%	74%
No_Title	100%	99.20%	99.60%	67.30%
No	99.60%	100%	99.60%	81.70%
Name_Title	94.30%	92.60%	90.50%	70.70%
Name	96%	90.90%	88.60%	75.30%
Date_Title	97.70%	96.20%	94.70%	74.90%
Date	99.90%	100%	99.60%	77%
Sex_Title	99.90%	100%	99.50%	75.30%
Sex	99.90%	100%	99.60%	76.80%
Nation_Title	99.50%	98.10%	98.50%	75.50%
Nation	100%	98.10%	98.80%	75.80%
Origin_Title	95.70%	96.20%	93.80%	73.20%
Origin	95.80%	89.30%	88.30%	68.50%
Residence_Title	97.70%	92.70%	91.60%	62.90%
Residence	98.60%	96.20%	96.50%	76.10%
Expiry_Title	99.70%	100%	99.50%	78.10%
Expiry	100%	100%	99.60%	75.40%

Table 2: Precision full-sentence comparison between pre-trained and transfer learning of VietOCR model.

Class	**Pre-Train Model**	**Transfer Learning Model**
All	64.23%	98.19%
ID_Title	73.82%	100%
ID	59.45%	97.65%
Name_Title	75.55%	100%
Name	49.83%	95.71%
DOB_Title	76%	100%
Date of birth	57.47%	96.12%
Sex_Title	74.66%	100%
Sex	57.44%	99.29%
Nation_Title	76.39%	100%
Nation	71.85%	100%
Origin_Title	70.06%	100%
Origin	50.24%	89.56%
Residence_Title	71.68%	100%
Residence	49.33%	94.07%
Expiry_Title	72.04%	100%
Expiry	53.17%	98.73%

Thanks to transfer learning to achieve high performance on the classification model (ResNet50) with an accuracy of 97.60 % on the test data. Figure 19 describes the training process. Overall, our implementation of the proposed method achieves good performance compared to the average performance of other research papers with an average system-wide accuracy of about 97.98 % with a run time of about 7 to 8 seconds including declared on the web application to inference as shown in Figure 20 .

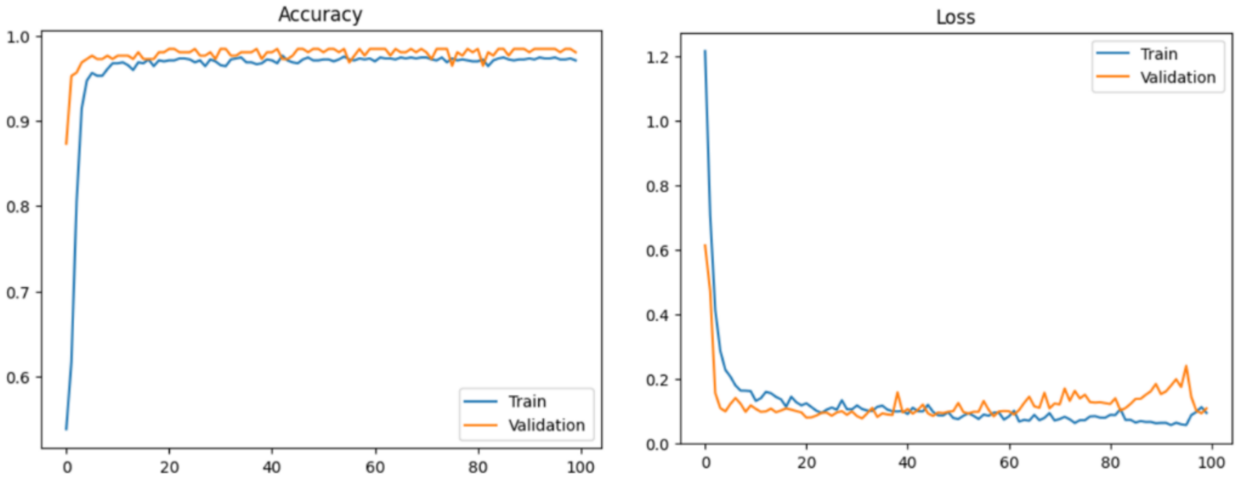
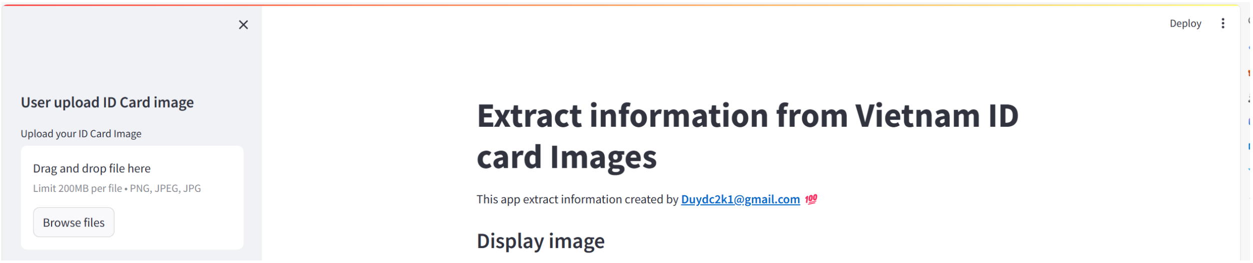


Figure 19: Accuracy and Loss on Classification model (Transfer learning ResNet50).



User upload ID Card image

Upload your ID Card Image

Drag and drop file here

Limit 200MB per file • PNG, JPEG, JPG

Browse files

Extract information from Vietnam ID card Images

This app extract information created by Duydc2k1@gmail.com 💯

Display image

Figure 20: Web-app interface.

5. CONCLUSION

We deliver significant improvements in efficiency and time accuracy, eliminating the need for manual and error-prone methods. We achieved high mean Average Precision (mAP) scores for ID card segmentation and text detection, with exceptional precision scores for full-sentence recognition. The efficient extraction of information from ID cards is crucial for various daily services in Vietnam. Our system has the potential to streamline and enhance these processes. The average accuracy of our system is influenced by both the background and lighting conditions of the image, thereby impacting the performance of the image segmentation model and the VietOCR transfer learning model in our proposed method.

In the future, our focus will be on improving the accuracy and efficiency of text detection and recognition in our system. Leveraging advancements in deep learning algorithms and techniques, we aim to enhance the system's ability to extract information accurately from various types of documents, including passports, invoices, and other common formats used in administrative and service processes.

References

[1] Chong Zhou. *Yolact++ Better Real-Time Instance Segmentation*. University of California, Davis, 2020.

[2] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[3] T. H. Chen. Do you know your customer? bank risk assessment based on machine learning. *Applied Soft Computing*, 86:105779, 2020.

[4] Hoan Tran Viet, Quang Hieu Dang, and Tuan Anh Vu. A robust end-to-end information extraction system for vietnamese identity cards. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 483–488. IEEE, 2019.

[5] Hoang Danh Liem, Nguyen Duc Minh, Nguyen Bao Trung, Hoang Tien Duc, Pham Hoang Hiep, Doan Viet Dung, and Dang Hoang Vu. Fvi: An end-to-end vietnamese identification card detection and recognition in images. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 338–340. IEEE, 2018.

[6] Nguyễn Hoàng Tú, Viên Thanh Nhã, Đỗ Thị Kim Dung, Tiếp Sỹ Minh Phụng, et al. Xây dỰng hỆthỐng trích xuẤt thông tin giẤy tỜ tuỲ thân tỪ hình Ảnh cho hỆthỐng ĐỊnh danh khách hàng ĐiỆn tỬ. 2022.

[7] H Emrah Tasli, Ronan Sicre, Theo Gevers, and A Aydin Alatan. Geometry-constrained spatial pyramid adaptation for image classification. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1051–1055. IEEE, 2014.

[8] Ahmad Montaser Awal, Nabil Ghanmi, Ronan Sicre, and Teddy Furon. Complex document classification and localization application on identity document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 426–431. IEEE, 2017.

[9] Thanh Cong Nguyen, Dinh Tuan Nguyen, and Quoc Long Tran. Information extraction from id card via computer vision techniques. Technical report, VNU University of Engineering and Technology, 2018.

[10] Peggy M. Andersen, Philip J. Hayes, Steven P. Weinstein, Alison K. Huettner, Linda M. Schmandt, and Irene B. Nirenburg. Automatic extraction of facts from press releases to generate news stories. In *Third Conference on Applied Natural Language Processing*, pages 170–177, Trento, Italy, March 1992. Association for Computational Linguistics.

[11] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[12] Wei-Ying Ma and BS Manjunath. Edge flow: a framework of boundary detection and image segmentation. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 744–749. IEEE, 1997.

[13] Harold N Gabow and Robert Endre Tarjan. A linear-time algorithm for a special case of disjoint set union. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 246–251, 1983.

[14] Michael B Dillencourt, Hanan Samet, and Markku Tamminen. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM (JACM)*, 39(2):253–280, 1992.

[15] Guizhong Fu, Peize Sun, Wenbin Zhu, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, and Yanpeng Cao. A deep-learning-based approach for fast and robust steel

surface defects classification. *Optics and Lasers in Engineering*, 121:397–405, 2019.

[16] Ibtissam Benchaji, Samira Douzi, Bouabid El Ouahidi, and Jaafar Jaafari. Enhanced credit card fraud detection based on attention mechanism and lstm deep model. *Journal of Big Data*, 8:1–21, 2021.

[17] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[18] Niloofar Tavakolian, Azadeh Nazemi, and Donal Fitzpatrick. Real-time information retrieval from identity cards. *arXiv preprint arXiv:2003.12103*, 2020.

[19] ChiatPin Tay, Vigneshwaran Subbaraju, and Thivya Kandappu. Privobfnet: A weakly supervised semantic segmentation model for data protection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2421–2431, 2024.

[20] Trinh Tan Dat, Nguyen Nhat Truong, Pham Cung Le Thien Vu, Vu Ngoc Thanh Sang, Pham Thi Vuong, et al. An improved crnn for vietnamese identity card information recognition. *Computer Systems Science & Engineering*, 40(2), 2022.

[21] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics*, 52(8):8574–8586, 2021.

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[23] Ross Girshick. Fast r-cnn in proceedings of the ieee international conference on computer vision (pp. 1440–1448). *Piscataway, NJ: IEEE.[Google Scholar]*, 2, 2015.

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[25] Deepika Gupta and Soumen Bag. Cnn-based multilingual handwritten numeral recognition: A fusion-free approach. *Expert Systems with Applications*, 165:113784, 2021.

[26] Duc Phan Van Hoai, Huu-Thanh Duong, and Vinh Truong Hoang. Text recognition for vietnamese identity card based on deep features network. *International Journal on Document Analysis and Recognition (IJDAR)*, 24:123–131, 2021.

[27] Trinh Tan Dat, Nguyen Nhat Truong, Pham Cung Le Thien Vu, Vu Ngoc Thanh Sang, Pham Thi Vuong, et al. An improved crnn for vietnamese identity card information recognition. *Computer Systems Science & Engineering*, 40(2), 2022.

[28] Guofeng Tong, Yong Li, Huashuai Gao, Huairong Chen, Hao Wang, and Xiang Yang. Ma-crnn: a multi-scale attention crnn for chinese text line recognition in natural scenes. *International Journal on Document Analysis and Recognition (IJDAR)*, 23:103–114, 2020.

[29] R Girshick, J Donahue, T Darrell, UC Berkeley, and J Malik. R-cnn: region-based convolutional neural networks. In *Proc. Comput. Vis. Pattern Recognit*, pages 2–9, 2014.

[30] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[31] Yiqing Zhang, Jun Chu, Lu Leng, and Jun Miao. Mask-refined r-cnn: A network for refining object details in instance segmentation. *Sensors*, 20(4):1010, 2020.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[35] Chien-Yao Wang, Hong-Yuan Mark Liao, and I-Hau Yeh. Designing network design strategies through gradient path analysis. *arXiv preprint arXiv:2211.04800*, 2022.

[36] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 984–993, 2018.

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[38] Chien-Yao Wang, Hong-Yuan Mark Liao, and I-Hau Yeh. Designing network design strategies through gradient path analysis. *arXiv preprint arXiv:2211.04800*, 2022.

[39] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 984–993, 2018.

[40] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. *arXiv preprint arXiv:2105.04206*, 2021.

[43] Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. *arXiv preprint arXiv:2306.04979*, 2023.