TU WIEN Informatics

Automatisierte Ausweisverifizierung mittels Deep Learning

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering & Internet Computing

eingereicht von

Simon Freitter, BSc.
Matrikelnummer 01633069

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Robert Sablatnig
Mitwirkung: Univ.Ass. Dipl.-Ing. Marco Peer, BSc

Wien, 28. Jänner 2025

Simon Freitter	Robert Sablatnig

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Automating ID Card Verification leveraging Deep Learning

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering & Internet Computing

by

Simon Freitter, BSc.
Registration Number 01633069

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Robert Sablatnig
Assistance: Univ.Ass. Dipl.-Ing. Marco Peer, BSc

Vienna, January 28, 2025

Simon Freitter	Robert Sablatnig

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Simon Freitter, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 28. Jänner 2025

Simon Freitter

Danksagung

An dieser Stelle möchte ich mich herzlich bei meinem Betreuer Marco Peer bedanken. Seine stets kompetente und zeitnahe Unterstützung hat maßgeblich zur Erstellung meiner Diplomarbeit beigetragen.

Ein großer Dank für ihre Unterstützung gilt auch meiner Familie, die mir die Möglichkeit gab, mich voll und ganz auf diese Arbeit zu konzentrieren.

Mein besonderer Dank gilt außerdem dem Roten Kreuz, das durch die Bereitstellung der Trainingsdaten entscheidend dazu beigetragen hat, dass diese Arbeit in ihrer vorliegenden Form realisiert werden konnte.

Acknowledgements

I would like to take this opportunity to thank my supervisor Marco Peer. His always competent and prompt support has contributed significantly to the preparation of my thesis.

I would also like to thank my family for their support, which gave me the opportunity to concentrate fully on this thesis.

My special thanks also go to the Red Cross, which made a decisive contribution to the realisation of this thesis in its present form by providing the training data.

Kurzfassung

In dieser Arbeit wird die Anwendung von Deep-Learning-Techniken für die automatische Überprüfung von Ausweiskarten untersucht, wobei der Schwerpunkt auf optischen Merkmalen liegt. Zu den wichtigsten Herausforderungen gehören die Handhabung begrenzter Trainingsdaten, Klassenungleichgewicht und unleserlicher Text aufgrund von Abnutzung oder Verpixelung. Das vorgeschlagene System besteht aus drei Kernmodulen: Segmentierung, Klassifizierung und Textextraktion von ID-Karten. Das fertige System wird anschließend vom Österreichischen Roten Kreuz eingesetzt.

Die Segmentierung erfolgt anhand des YOLOv8-Modells, das die ID-Karte im Eingabebild identifiziert. Die Leistung des Modells wird anhand der in dieser Studie eingeführten Metrik Recall@IoU-MinIS bewertet. Die vom YOLOv8-Modell ausgewählte Bildregion wird anschließend als Eingabe für das Klassifizierungsmodell verwendet. Für die Klassifizierung werden mehrere Ansätze mit einem ResNet-Backbone bewertet: Cross-Entropy Loss, Triplet Margin Loss und Angular Margin Loss. Für die Textextraktion werden drei OCR-Technologien bewertet: Tesseract, EasyOCR und PaddleOCR. Darüber hinaus wird die mögliche Leistungssteigerung durch die zusätzliche Anwendung des multimodalen Large Language Models Llama 3.2 Vision untersucht.

Bei Anwendung der Recall@IoU-MinIS-Metrik erreicht das YOLOv8-Modell eine genaue Segmentierung für 90% der ID-Karten. Innerhalb des Klassifizierungsmoduls erreichen Modelle, die auf Angular Margin basierendem Metric Learning beruhen, eine maximale Genauigkeit von 98,07% und eine normalisierte Genauigkeit von 97,08% und übertreffen damit die anderen Ansätze. Die Robustheitsanalyse verdeutlicht die Herausforderungen bei der Unterscheidung visuell ähnlicher ID-Karten, auf denen das Modell nicht trainiert wurde. Dieses Problem wird durch die Einbeziehung von Konfidenzschwellen gemildert. Bei der Textextraktion zeigt PaddleOCR eine überlegene Leistung, indem es 93,35% aller Felder korrekt extrahiert und bei 82,51% der ID-Karten vollständige Korrektheit erreicht. In Kombination mit EasyOCR verbessern sich diese Werte auf 95,30% bzw. 87,43%. Die Integration von Llama 3.2 Vision steigert die Genauigkeit der vollständigen Ausweiserfassung weiter auf 96,72%, allerdings mit erheblich höherem Rechenaufwand. Mit einer Toleranz für die Fehlerspanne extrahiert PaddleOCR unabhängig 97,89% aller Felder korrekt und erreicht eine vollständige Korrektheit für 93,99% der Ausweise.

Abstract

This thesis examines the application of deep learning techniques for automated ID card verification, focusing on optical characteristics. The key challenges addressed include managing limited training data, class imbalance, and text obscured by wear or pixelation. The proposed system consists of three core modules: ID card segmentation, classification, and text extraction. The completed system is subsequently used by the Austrian Red Cross.

For segmentation, the YOLOv8 model identifies the ID card within the input image. The model's performance is evaluated using the Recall@IoU-MinIS metric introduced in this study. The image region selected by the YOLOv8 model is subsequently utilized as input for the classification model. For classification, multiple approaches using a ResNet backbone are assessed, including Cross-Entropy Loss, Triplet Margin Loss, and Angular Margin Loss. For text extraction, three OCR technologies are evaluated: Tesseract, EasyOCR, and PaddleOCR. Furthermore, the potential enhancement of performance through the auxiliary application of the multimodal large language model Llama 3.2 Vision is investigated.

Applying the Recall@IoU-MinIS metric, the YOLOv8 model achieves accurate segmentation for 90% of the ID cards. Within the classification module, models utilizing angular margin-based metric learning attain a maximum accuracy of 98.07% and a normalized accuracy of 97.08%, outperforming other approaches. Robustness analysis highlights challenges in differentiating visually similar ID cards, which the model has not been trained on. This issue is mitigated by incorporating confidence thresholds. For the text extraction task, PaddleOCR demonstrates superior performance, accurately extracting 93.35% of all fields and achieving full correctness for 82.51% of ID cards. When combined with EasyOCR, these metrics improve to 95.30% and 87.43%, respectively. The integration of Llama 3.2 Vision further increases the accuracy of complete ID card extraction to 96.72%, though at a significantly higher computational cost. Allowing for an error margin tolerance, PaddleOCR independently extracts 97.89% of all fields correctly and achieves complete correctness for 93.99% of ID cards.

Contents

CHAPTER 1

Introduction

Neural Networks (NNs) have been integral to numerous technological advancements and are now embedded in various aspects of everyday life [LBH15], often without users' awareness. These versatile models are employed across a wide range of tasks, utilizing different architectures and learning paradigms to achieve their goals. [Sch15] Their adaptability and effectiveness have made them indispensable tools in fields such as image and speech recognition, natural language processing, and content recommendation. [LBH15] One potential application of neural networks is automated ID card verification. Although there exist scientific works, that address neural network-powered ID card verification [CPIS19, LVS+21, GSRK21, WXW+19], they primarily focus on comparing the image on the card to the individual's face for authentication or extracting textual information from the ID card. In most cases, these studies have access to large datasets for model training. To the best of my knowledge, no scientific work exists that focuses on evaluating ID cards, explicitly addressing the identification of their optical characteristics. However, manually verifying ID cards is a tedious and time-consuming task that requires significant attention to detail. An IT-driven solution can automatically determine the authenticity of an ID card by analyzing its optical characteristics and extracting the textual data. Convolutional Neural Networks (CNNs), a powerful class of deep learning models, are extensively utilized for the classification of objects across various domains [LBH15], making them well-suited for this task.

The aim of this thesis is to investigate and establish the extent to which deep learning methodologies can be effectively utilized for the automatic verification of identification cards based on their optical characteristics. Specifically, this research seeks to determine the critical factors and methodologies that influence the performance in ID card verification. This includes an in-depth examination of image quality and quantity, assessing how factors such as resolution, lighting, clarity, and the number of training images affect system performance. Additionally, the thesis investigates various deep learning approaches to determine the most suitable one for this task.

The dataset comprising 183 positive instances of ID card images used for training and evaluating the model is supplied by the Austrian Red Cross. These instances are distributed unevenly across 11 distinct categories, representing 8 Austrian states and 3 subcategories for the state of Tirol. A significant challenge in accurately verifying these ID cards arises from the visual variability of ID cards across different Austrian states. Therefore, it is important to investigate whether a model trained with all valid types of ID cards grouped under a single class or with each valid type assigned to an individual class would be more appropriate for this task. This exploration is particularly crucial given the significantly lower number of sample ID card images available for this study compared to other studies.

For verifying the ID cards, a pipeline comprising three modules as shown in Figure 1.1), which are described in detail in Chapter 3, is employed. The input image is initially processed by a model employing the YOLO object detection algorithm to identify the Region of Interest (ROI), specifically a check card. The extracted check card is subsequently processed by the classification model, where it is categorized as either valid or invalid. Based on the model's training, the output may be a straightforward "valid" or "invalid" classification, or it may include a specific designation of the valid category to which the ID card belongs. If classified as valid (or as belonging to one of the valid categories), the image is subsequently processed by the Optical Character Recognition (OCR) model to extract the personal information from the ID card.

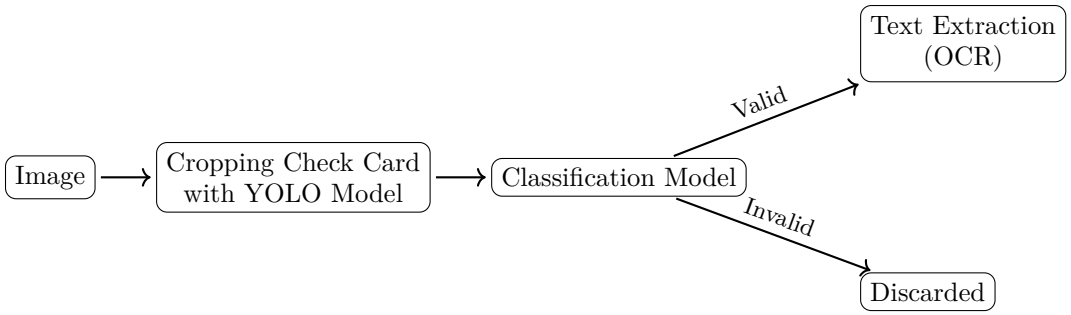


Figure 1.1: Process of verifying the ID card from image input to classification and potential text extraction or discarding.

The proposed verification application is utilized within the webshop of the Lower Austrian State Association of the Red Cross to authenticate members of all Austrian State Associations of the Red Cross. Consequently, the application must facilitate real-time authentication.

Due to the sensitive nature of the data and the necessity to extract personal information without anonymization, in accordance with an agreement from the Austrian Red Cross, the images will be exclusively stored and processed on a secure local machine. Note: **All rights to the designs of the ID cards are reserved by the Red Cross.**

In summary, this thesis aims to address the following research questions:

> To what extent can deep learning methodologies be leveraged for the automatic verification of ID cards based on their optical characteristics? These characteristics include background patterns or images, logos, and decorative elements such as border lines.
>
> Which contemporary OCR methodology is most effective for extracting textual data from ID cards?

This research does not authenticate individuals by matching their facial features to the photograph on the ID card. Instead, the primary focus is on verifying the authenticity of the ID card itself, with a secondary objective of extracting the ID card's textual data.

The thesis is structured the following way: Chapter 2 gives a review of existing research and technologies in related fields. Chapter 3 elaborates on the proposed pipeline, including data preparation, ID card segmentation, classification, and text extraction techniques. Chapter 4 assesses the performance of these methods across various metrics, explores robustness, and identifies key challenges. In Chapter 5, the thesis concludes with a summary of findings and suggestions for future work.

CHAPTER 2

Related Work

As presented in Chapter 1, the proposed pipeline comprises three distinct tasks: ID card segmentation, ID card classification, and text extraction from the ID card. In the domains of ID card segmentation and text extraction, prior research has focused on automating the processing of ID cards. Moreover, segmentation and text extraction tasks are critical in various other contexts as well.

2.1 Segmentation

A widely utilized framework in the domain of object detection is YOLO [RDGF16]. YOLO addresses object detection by framing it as a single regression problem, wherein bounding boxes and class probabilities are predicted from entire images in one unified pass. This approach enables YOLO to process images at a rate of 45 frames per second, and a more compact variant, Fast YOLO, achieves processing speeds of up to 155 frames per second, thereby enabling real-time image processing capabilities. Over the years, several versions of YOLO have been developed, with the latest officially published version being YOLOv7 [WBL23]. The most recent preprint version available is YOLOv10 [WCL+24].

Another methodology for object detection involves employing the Transformer architecture, as demonstrated in DETR (DEtection TRansformer) [CMS+20]. DETR reconceptualizes object detection as a direct set prediction task, thereby streamlining the process by eliminating the necessity for manually crafted components. Through the utilization of a fixed set of learned object queries, DETR efficiently interprets object relationships and the overall image context, enabling concurrent prediction generation.

Lara et al. [LVS+21] explored the segmentation of ID cards from their backgrounds as a fundamental step towards extracting relevant information from images of ID cards. In their study, they refined the input images by modifying them, such as isolating the ID card from its surroundings and introducing varied backgrounds or adjusting color schemes,

to enhance the precision of the models. Notably, their experiments demonstrated that a MobileUNet architecture yielded the most favorable outcomes. However, they also found that DenseNet10 [VAT20], a lightweight version of DenseNet, comprising 210,732 parameters (compared to 6.5 million parameters in MobileUNet), delivered competitive performance in this context.

2.2 Text Extraction

Gupta et al. [GSRK21] investigate the automation of textual data extraction from ID cards to accelerate authentication procedures and enhance extraction accuracy. Leveraging the OpenNMT architecture, their developed OCR model for ID numbers showed enhanced performance compared to the Tesseract OCR engine across a test set comprising 36,000 ID numbers.

Wu et al. [WXW+19] propose a comprehensive framework designed to extract textual data from identification cards and authenticate individuals by matching their faces to the photographs on the ID cards. The framework comprises two primary components: a face verification model named Inception-ResNet Face Embedding (IRFE) and a text extraction method referred to as Morphology Transformed Feature Mapping (MTFM). Empirical evaluations indicate that the proposed framework outperforms existing state-of-the-art methodologies in both face verification accuracy and text recognition from ID cards.

2.3 Integration of Segmentation and Text Extraction in other Domains

The study conducted by Praneeth Reddy et al. [RSHS24] investigates the application of YOLOv8 for license plate detection and assesses the performance of three OCR models – EasyOCR, PaddleOCR, and Tesseract – based on their confidence scores. The methodology includes detecting and segmenting license plates using YOLOv8, followed by image preprocessing steps such as grayscale conversion, noise reduction, and morphological filtering. Subsequently, the OCR models are employed for text extraction. The results indicate that PaddleOCR consistently achieves the highest confidence scores, demonstrating its reliability in character recognition, whereas EasyOCR and Tesseract exhibit greater variability in performance.

CHAPTER 3

Methodology

In this chapter, the methodology for implementing deep learning-based ID card verification is outlined. Section 3.1 describes the process of dataset creation and the necessary preprocessing steps. Section 3.2 details the segmentation procedure employed to isolate ID cards from background elements, utilizing the YOLO framework. Furthermore, a reliable performance metric tailored to this specific scenario is proposed. Section 3.3 elaborates on the classification techniques applied, including traditional cross-entropy-based methods and advanced metric learning approaches, such as Triplet Margin Loss and ArcFace. While the implemented methods are designed to classify ID cards into 12 distinct types, thereby assigning valid ID cards to their respective categories, potential benefits of a simpler 2-class classification approach are also discussed. Lastly, Section 3.4 explains the text extraction process leveraging various OCR techniques and examines the potential of multimodal large language models to enhance performance. Additionally, strategies for improving robustness through fault-tolerant mechanisms are presented.

3.1 Dataset Construction and Preprocessing

This section describes the creation and preprocessing of the dataset. It specifically addresses the variability in the design and condition of ID cards, as well as the measures taken to ensure robustness against a wide range of negative instances. The preprocessing steps are structured to optimize the effectiveness of the subsequent model training.

3.1.1 Dataset Creation

The dataset used in this research consists of positive instances, which represent ID cards issued by the Austrian Red Cross, and negative instances, which include invalid or irrelevant check cards that are not necessarily ID cards. For training both the segmentation and classification models, 15% of the samples are allocated for training, while 85% are used for evaluation.

Positive Instance Collection

The positive samples, provided by the Austrian Red Cross, consist of identification cards from the nine different states of Austria (see Table 3.1). A key challenge stems from the significant variation in design and features across these cards (see Figure 3.1). Additionally, the state of Tirol issues three distinct types of ID cards for different purposes, all of which are considered valid, thus increasing the total to eleven distinct categories of valid ID cards. The designs of the ID cards share only the logo, which includes the name of the corresponding state. Although the state's name differs, the design –characterized by white text on a gray background – provides a consistent visual appearance for the names across various states, with the exception of ID cards belonging to the "Vienna" class. However, apart from this logo, there are no uniform visual characteristics or standardized textual elements (such as name, date of birth, or validity date) present across all ID cards.

ID Category	Number of available samples
Burgenland	8
Kaernten	13
Niederoesterreich	10
Oberoesterreich	63
Salzburg	4
Steiermark	10
Tirol-Bestaetigungsausweis	2
Tirol-Rettungsdienst	53
Tirol-Sanitaeter	12
Vorarlberg	5
Wien	3
Invalid check cards	61

Table 3.1: Distribution of available samples by ID category across Austrian regions and specific ID types, including invalid check cards. The state of Tirol has multiple valid ID cards with distinct optical appearances, necessitating further subdivision. A significant imbalance is observed in the number of available samples across the different categories.

Negative Instance Collection

The negative instances consist of a selection of check and ID cards that generally do not share common characteristics (see Figure 3.2). The purpose of these negative instances

(a) Burgenland

(b) Kaernten

(c) Niederoesterreich (NOE)

(d) Oberoesterreich (OOE)

(e) Salzburg

(f) Steiermark

(g) Tirol (Bestaetigung)

(h) Tirol (Rettungsdienst)

(i) Tirol (Sanitaeter)

(j) Vorarlberg

(k) Wien

Figure 3.1: Examples of each category of valid ID cards: The Red Cross logo is the only consistent visual element across all card types. Also, there is no standardized set of textual elements (e.g., name, date of birth, or validity period) shared between different types of ID cards. Note: The ID cards shown here were selected for their pristine condition, free from visible soiling or abrasion. The dataset, however, also includes images of ID cards that are partially covered by fingers or ID cases, as well as cards that show signs of wear and soiling.

is to ensure the model's robustness against invalid IDs or other check cards[1], which the model might otherwise incorrectly classify as valid ID cards. However, the negative instances do not include forged IDs or other cases that explicitly attempt to mimic valid ID cards.

Figure 3.2: Examples from the dataset include images of invalid ID cards and various check cards, each segmented from their backgrounds. In addition to the displayed images, the dataset includes a wide range of other ID cards as well as various non-ID cards, such as debit cards, membership cards, and miscellaneous check cards. The diversity in appearance ensures the model learns to distinguish valid ID cards from other types of cards effectively.

3.1.2 Data Preparation

The collected dataset, consisting of both positive and negative instances of ID card images, is subjected to preprocessing and formatting to prepare it for the training and evaluation stages of the neural network. The first step is to annotate the data. Since the initial phase of the verification process involves detecting the ID card within an image using the YOLO framework, the data is annotated in the format required by YOLO (‘<class> <x-center> <y-center> <width> <height>‘).

To ensure the classification model accurately learns the relevant patterns, each ID card image is cropped according to the ground truth bounding box provided by the YOLO annotations. Additionally, the images are rotated so that the right side of the ID card is positioned at the top, standardizing their orientation for the model.

[1]In the context of this thesis, any check or ID card that is not a Red Cross member card is considered invalid and is used as a negative instance. This does not imply that the card may not be valid in other contexts.

3.2 ID Card Segmentation using YOLO

The initial phase involves training a model specifically for detecting check cards. This step is necessary for isolating ID cards from background elements, ensuring that the classification model focuses solely on the ID card features without interference from irrelevant contextual information. Additionally, segmentation addresses the challenge of image preprocessing for classification, where resizing images to a uniform shape can introduce significant distortion or pixelation, potentially compromising recognition accuracy. Due to its precise yet computationally efficient nature, the YOLO framework, specifically YOLOv8 [JCQ23], is employed for this purpose.

3.2.1 Metrics

To evaluate the performance of the YOLO model on this task, the following two questions are answered:

1. How many of the ID/check cards are detected?
2. How precise is the cropping of the detected ID/check cards?

Key performance metrics for evaluating a YOLO model include Intersection over Union (IoU), Precision, Recall, (Mean) Average Precision (AP/mAP), and the F1 Score [Ult23]. Precision measures the proportion of correctly identified check cards relative to all items classified as check cards by the YOLO model.[2] However, false positives (i.e., instances where non-check card objects are mistakenly detected as check cards) are of negligible relevance in this specific application. This is because any region detected as a check card will subsequently be evaluated by a classification model, which is responsible for identifying invalid ID cards.

False positives, however, lead to greater computational demands at the classification stage. In extreme cases, a significant rise in false positives could potentially result in a denial of service, though this remains a largely theoretical risk.

Due to the negligible practical importance of Precision in this context, metrics that depend on it – such as (Mean) Average Precision and the F1 Score – are also not used for assessment.

The detection rate of ID/check cards is quantified using the Recall metric, as it indicates how many of the relevant cards are detected. However, Recall alone does not provide any insight into the accuracy of the cropping of the detected ID/check cards. To address this limitation, the inclusion of the IoU metric is essential. IoU evaluates the overlap between the predicted and ground truth bounding boxes, offering a measure of the precision of

[2] Precision in this context is not related to the issue of how precise a check card is cropped, but how many of the detected areas are actually ID cards.

the cropping. This combined metric, referred to as Recall at a specific IoU threshold (**Recall@IoU**), provides a more informative model evaluation measure.

Mathematically, Recall and IoU are defined as follows:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Recall@IoU with Minimum Intersection Score

However, including the IoU, the Recall@IoU score alone is not suitable for assessing the YOLO model's performance in this context. This is because the evaluation depends not only on the IoU score itself, but also on the individual values of the intersection and union areas, which contribute to the final IoU value. Since including some background within the detected region is less problematic than excluding parts of the ID card, in this context, a high intersection score is more critical than a high union score. For example, an IoU score of 0.8 is acceptable if the deviation from perfect crop comes from including additional 25% of background (see Figure 3.3), as this does not hinder classification. However, missing 20% of the ID card would lead to discarding a valid card or excluding essential text.

Moreover, the Intersection score alone cannot fully capture model performance, as a 100% intersection is achieved by detecting the entire image as an ID card. To address this issue, Recall@IoU is extended to incorporate a Minimum Intersection criterion and introduce a new metric specifically for evaluating the model's performance in this use case: **Recall@IoU with Minimum Intersection Score** (**R@IoU-MinIS**).

Formally, R@IoU-MinIS is defined as:

$$\text{R@IoU-MinIS} = \frac{|\{\text{TP}_i \mid \text{IoU}(\text{TP}_i, \text{GT}_i) \geq T_{\text{IoU}} \wedge \text{IS}(\text{TP}_i, \text{GT}_i) \geq T_{\text{IS}}\}|}{N}$$

where:

- TP_i is the true positive prediction of the object i[3],
- GT_i is the corresponding ground truth of the object i (i. e. a check card),
- $\text{IoU}(\text{TP}_i, \text{GT}_i)$ is the Intersection over Union Score between TP_i and GT_i,
- $\text{IS}(\text{TP}_i, \text{GT}_i)$ is the Intersection Score between TP_i and GT_i,

[3]If multiple true positive predictions exist for the same check card, the prediction with the highest IoU is selected for evaluation.

- T_{IoU} denotes the IoU threshold,
- T_{IS} denotes the Intersection threshold,
- N denotes the total number of detectable instances.

Figure 3.3: An ID card with an Intersection over Union (IoU) score of 0.8, illustrating that a moderate IoU value does not inherently indicate improper cropping or the omission of critical sections of the ID card. The green boundary represents the area predicted by the YOLO model, while the blue boundary along the edges denotes the ground truth.

3.3 ID Card Classification

The classification module depicts the main part of this thesis. Motivated by its capability to detect objects in images, in this thesis YOLO's capability of directly assigning ID cards to their respective class is examined. However, due to its poor results in this task (as discussed in Section 4.2.1, YOLO's deployment for the proposed pipeline is limited to detection and segmentation of check cards within images.

The following approaches utilize a ResNet architecture pre-trained on the IMAGENET1K_V1 dataset, focusing on categorizing images into 12 distinct classes (11 valid classes and 1 invalid class).

For the ResNet-based classification model, additional preprocessing is applied: Firstly, the images need to be reshaped to ensure compatibility with the model's input requirements. Secondly, the `ColorJitter` transformation was applied for image augmentation by adjusting brightness, contrast, and saturation by ± 0.2 and hue by ± 0.1. Also, due to the imbalance in the number of images per class (e.g., 2 samples for class "Tirol-Bestaetigungsausweis" while 63 for class "Oberoesterreich"), a data sampler is employed to address this issue and ensure a more balanced training.

Due to the imbalance of the dataset, the normalized accuracy metric is examined in addition to standard accuracy. Standard accuracy is defined as the ratio of correctly classified ID/check cards to the total number of samples. In contrast, normalized accuracy is calculated as the accuracy of each class divided by the number of classes. This metric ensures that each class contributes equally to the evaluation metric, regardless of class sample size.

To make the results comparable, additionally to fixating the training and validation data split, the following parameters were fixed:

Optimizer	Batch size	Number of epochs
Adam [Kin14] with a learning rate of 0.0001	16	50

3.3.1 Approach with Cross-Entropy Loss

The first approach incorporates the Cross-Entropy Loss (CEL), which serves as a fundamental method for classification tasks by comparing the predicted probability distribution with the true labels. It is defined as follows:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where:

- y_i is the ground truth label for class i (a one-hot encoded vector).
- $\hat{y}_i$ denotes the predicted probability for class i, as produced by the model.
- N corresponds to the total number of classes involved in the classification task.

3.3.2 Metric Learning Approach with Triplet Margin Loss

The second approach to address the classification task involves utilizing a deep metric learning method [KB19] to cluster the ID card features in the embedding space. The triplet margin loss [SKP15] is chosen as the loss function, which is defined as follows:

$$L = \sum_{i=1}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

where:

- $f(x)$ is an embedding function, in this case the NN, that maps the input x to a feature space.

- x_i^a, x_i^p, and x_i^n represent the anchor, positive, and negative samples, respectively, for the i-th triplet.
- α is the margin, which enforces a minimum separation between the anchor-positive and anchor-negative distances.
- N is the total number of triplets in the batch.

Since several classes possess only a single training sample, a k-Nearest Neighbors (k-NN) classifier is employed with $k = 1$. Based on promising results in multi-class classification tasks, ResNet18 and ResNet34 architectures are used with input image sizes of 64×64, 112×112 or 224×224 pixels.

3.3.3 Angular Margin-Based Metric Learning Approach

Motivated by its performance for the task of face recognition, this study explores the integration of ArcFace [DGXZ19], an angular margin-based metric learning technique, into the model. ArcFace introduces a way of discriminative learning by adding an angular margin penalty between the learned feature embeddings and their corresponding class centers in the hypersphere space in the following way:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)}+\sum_{j\neq y_i}e^{s\cdot\cos(\theta_j)}}$$

where:

- θ represents the angle between the weight vector $\mathbf{w}_j$ of the j-th class and the feature vector $\mathbf{x}_i$ of the i-th sample.
- y_i denotes the ground-truth class label of the i-th sample.
- m is an angular margin added to θ_{y_i}, encouraging larger angular separation between classes.
- s is a scaling factor applied to control the magnitude of the logits before the softmax computation.
- N represents the total number of samples in the current training batch.

ArcFace is integrated as the final layer of the ResNet architecture, replacing the standard fully connected classification layer. The ArcFace loss is employed with an angular margin of m=0.3 and scale s=15.

3.3.4 12-Class Approach vs 2-Class Approach

Given that all distinct classes share a common optical characteristic in the form of a logo, the thesis incorporates experiments that consider only two classes: valid and invalid cards. The study evaluates whether this binary classification approach enhances accuracy by addressing the challenge of a limited sample size for multiple classes and investigates the potential disadvantages associated with this method.

3.3.5 Assessing Robustness of the Approach through Class Omission

Training the model on different types of valid ID cards evaluates its ability to distinguish between various valid ID card types while rejecting unrelated or arbitrary ID/check cards that do not closely resemble Red Cross ID cards. However, this approach does not provide information on the model's performance when presented with ID cards that exhibit a higher degree of similarity, such as Red Cross ID cards from other countries. To address this limitation, the study simulates a scenario involving visually similar ID cards as follows: First, the models are trained on all valid ID card classes except for one. Then, the excluded class is used during evaluation to determine whether the model correctly classifies these ID cards as invalid or incorrectly assigns them to one of the other valid classes. For reliable evaluation, an angular margin-based ResNet18 model was employed, trained on images with a resolution of 64×64 pixels, as this combination yielded the best performance during preliminary evaluations.

Incorporating Confidence Measures in Predictions

This thesis also examines methods to enhance the model's robustness against ID cards that display optical characteristics similar to those of valid ID cards. Rather than merely outputting the predicted class for an ID/check card, the model applies the softmax function to generate an output vector containing the probability distribution across all classes. The highest probability in this vector can be interpreted as the confidence of the model's prediction.

A confidence threshold involves a balance between two competing outcomes: the proportion of invalid ID cards correctly discarded because their confidence scores fall below the threshold (True Negatives) and the proportion of valid ID cards mistakenly discarded for the same reason (False Negatives). This thesis analyzes which confidence threshold achieves the optimal ratio of True Negatives to False Negatives.

3.4 Text Extraction via OCR

The final stage of the ID card verification pipeline involves text extraction to compare the data from the individual's ID card with the information provided. However, as illustrated in Table 3.2, the data available on each ID card type varies. In the proposed pipeline, Name, Date of Birth and Service Number are matched to provided information while the Validity Date is compared to the current date.

ID Category	Name	Validity Date	Date of Birth	Service Number
Burgenland	✓	✗	✓	✓
Kaernten	✓	✗	✗	✓
Niederoesterreich	✓	✓	✓	✓
Oberoesterreich	✓	✗	✓	✓
Salzburg	✓	✗	✗	✓
Steiermark	✓	✓	✓	✓
Tirol-Bestaetigung	✓	✗	✓	✗
Tirol-Rettungsdienst	✗	✓	✗	✓
Tirol-Sanitaeter	✓	✓	✓	✓
Vorarlberg	✓	✓	✓	✓
Wien	✓	✓	✗	✓

Table 3.2: The presence of varying textual information across different categories of ID cards requires identifying the specific type of ID card being verified. This is essential during the stage where the extracted text is matched with the provided reference text, as it determines which specific information must be present for accurate verification.

The suitability of the OCR module for this use case is determined by its ability to accurately extract the relevant fields while maintaining a fast processing time. Therefore, the models are evaluated using three specific metrics. First, their average execution time per image is analyzed. Additionally, the models are evaluated based on the proportion of fields correctly extracted. In this scenario, the text extraction is considered successful only if all relevant fields are correctly extracted. Therefore, the evaluation includes the proportion of ID cards for which all required fields are accurately captured. Based on these criteria, the performance of three OCR engines – Tesseract, EasyOCR, and PaddleOCR – is analyzed. Furthermore, this thesis explores whether the overall performance can be improved by combining the outputs of these different OCR engines.

Improving Accuracy with Contextual Understanding

The task of text extraction is not limited to traditional OCR technologies. Multimodal Large Language Models (MLLMs) demonstrate significant potential in tasks such as image analysis and have shown promising results in the domain of text recognition [LLH+24]. In this context, before evaluating metrics such as accuracy and speed, the selection of MLLMs is constrained by the requirement that the model must function locally to preserve privacy. One example of an MLLM that meets this criterion is "Llama 3.2

Vision". Therefore, the applicability of the Llama model in this scenario is investigated.

Since MLLMs are not explicitly designed for OCR tasks, textual instructions to guide the model in extracting text are provided. For this analysis, the exact instruction provided to the model is:

"Lies sämtlichen Text aus, den du auf diesem Bild lesen kannst. Als Antwort gib ausschließlich den Text auf diesem Bild wieder. Gib exakt den Text wieder, ohne zusätzliche Kommentare oder Erklärungen."

(Translation: "Read all the text that you can see in this image. As a response, provide only the text from this image. Reproduce the text exactly, without additional comments or explanations.")

Enhancing Performance Through Fault Tolerance

As the relevant fields include recurring patterns, such as variations in date formats and common errors like German umlauts, it is necessary to denominate the pattern and quantify deviations between erroneous detections and the ground truth. For quantification, the Levenshtein distance is utilized. The Levenshtein distance measures the minimum number of single-character edits – insertions, deletions, or substitutions – required to transform one string into another. This provides a metric for assessing the magnitude of the error. Consequently, after text extraction, the investigation focuses on whether recurring error patterns can be mitigated by introducing a certain degree of fault tolerance.

CHAPTER 4

Evaluation

In this chapter, the performance of the components within the proposed ID card verification pipeline is systematically evaluated. Section 4.1 focuses on the evaluation of the YOLO-based segmentation methodology. Section 4.2 presents a comparative analysis of various classification strategies, including cross-entropy-based models, metric learning using Triplet Margin Loss, and Angular Margin-Based Metric Learning (ArcFace). Additionally, it contrasts the 12-class categorization framework with a simplified 2-class (valid–invalid) classification scheme. Grad-CAM is employed to visually interpret the inference results for both the 12-class and 2-class approaches. The robustness of the system is also tested by simulating ID cards with similar optical characteristics. Section 4.3 evaluates text extraction methods, such as Tesseract, EasyOCR, and PaddleOCR, and explores the potential of MLLMs to enhance recognition accuracy. Furthermore, the applicability of incorporating an error margin tolerance is discussed as a strategy to improve the performance of text extraction.

4.1 ID Card Segmentation using YOLOv8

To evaluate the performance of YOLO model tasked with segmenting the ID cards from their background, the proposed R@IoU-MinIS score is used. Specifically, the IoU threshold is set to 0.8, and the Minimum Intersection Score (MinIS) is set to 0.9. Based on these parameters, the model with the highest performance, as determined by the YOLO framework, achieves a **R@IoU$_{0.8}$-MinIS$_{0.9}$ score of 90%**.

Although the R@IoU-MinIS score provides a measure of how accurately the ID card has been cropped, the R@IoU-MinIS is dependent on individual instances to determine whether critical parts of the ID card have been excluded. For skewed images, a lower Intersection score may emerge, even when only portions of the corners are excluded. This exclusion inherently also affects parts of the background, while the critical regions of the ID card could remain fully contained within the designated area of interest (see Figure

4.1). This phenomenon arises from the axis-aligned cropping approach employed by the selected YOLO model. The proportion of ID card content relative to background that is removed depends on the degree of skew in the ID card.

To accurately evaluate how well the R@IoU-MinIS score reflects the performance of the model, the individual crop outputs produced by the YOLO model are examined. The analysis reveals that only 0.97% of the ID card crops with an IoU score below the specified threshold still include all relevant parts. Further, the percentage of ID card crops with an IoU score above the threshold that omit relevant parts is also 0.97%. Consequently, the R@IoU-MinIS score gives a robust approximation of the proportion of ID cards that are accurately detected.

To achieve a higher degree of precision in evaluation, it would be necessary to consider additional parameters – such as the angular distortion of the ID card – or to design a more comprehensive metric that directly compares the classification outcomes of the human-cropped ID card with those of the YOLO-cropped counterpart. However, this approach depends on the behavior and performance of the classification model employed.

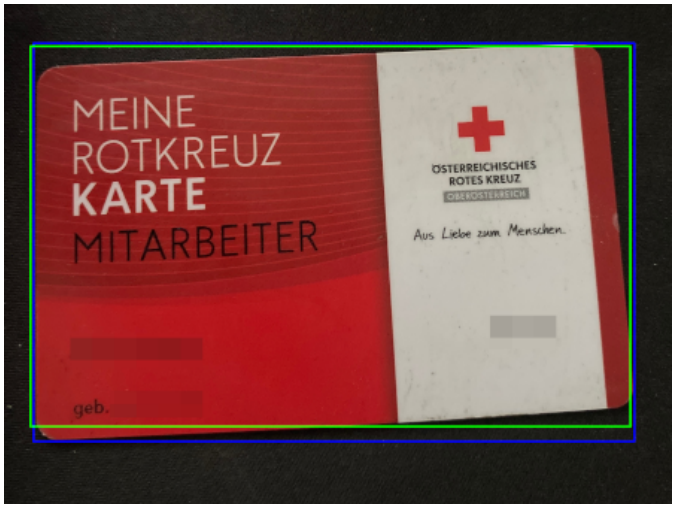


Figure 4.1: For images with skewed orientations, the R@IoU-MinIS score can be misleading. In this example, the prediction achieves an IoU and Intersection score of less than 0.95, despite only a small portion of the ID card's corner being removed. The skew causes the actual affected area of the ID card to be smaller than 5%, as the ground truth includes a certain amount of background.

4.1.1 Primary Detection Issue Patterns of the YOLO Model

In this section, the recurring patterns among the wrongly detected ID cards are investigated. An analysis of individual predictions made by the YOLO model reveals two recurring patterns among the improperly cropped check cards:

1. **Scanned ID Cards**: The model encounters difficulties with scanned ID cards that have plain, white backgrounds, especially when there is insufficient contrast between the edges of the ID card and the background (see Figure 4.2).

2. **ID Cards with Hard Edges in the Design**: The presence of distinct, straight edges with high color contrast within the design of the ID card causes inadequate cropping. These design features mislead the model, causing it to crop the ID card incorrectly along these edges or to detect multiple, separate instances of the same ID card (see Figure 4.3).

Figure 4.2: The YOLO model encounters challenges in detecting ID cards when there is no clearly visible boundary between the ID card and the background. This issue leads to a complete failure in detecting the ID card, as demonstrated in this example.

(a) (b) (c)

Figure 4.3: ID Cards with hard edges leads to the model either detecting several check cards within one check card (4.3a & 4.3b) or cutting off parts of the check card (4.3c)

4.2 ID Card Classification

In this section, the following approaches for ID card classification are evaluated: It examines YOLOv8 for ID card classification, a classification approach with Cross-Entropy

Loss, metric learning with Triplet Margin Loss, and Angular Margin-Based Metric Learning (ArcFace). To minimize the influence of random variation in individual results, each ResNet-based model is trained three times (unless specified otherwise), and the best outcome is selected for analysis. Detailed results are provided in the appendix.

4.2.1 YOLOv8 Classification Approach

The results obtained in Section 4.1 raise the question of whether YOLO can be effectively utilized to classify ID cards into appropriate categories or mark them as invalid. For this classification task, the evaluation metrics differ from those applied in assessing the image cropping model, as for this task, the primary objective is to determine the validity of the ID card. Consequently, precision becomes a critical metric, while the IoU metric is less relevant, provided that only classification accuracy is considered and not text extraction.

However, as shown in Figure 4.4, even at an IoU threshold of 0.5 − the standard value at which YOLOv8 computes its performance metrics ¯ the YOLO model demonstrates poor results with a best mAP@50[1] of 0.31.

4.2.2 Approach with Cross-Entropy Loss

In the initial approach, the ID card verification task is approached as a multi-class classification problem leveraging Cross-Entropy Loss. The best result achieves an accuracy of 95.17% and a normalized accuracy of 93.25% (see Table 4.1). Under these conditions, a slight trend emerged: reducing model size and image resolution correlated with an increase in accuracy. This trend is particularly noticeable in the normalized accuracy, suggesting that larger models and higher resolutions may struggle to capture relevant visual features in classes with limited sample sizes.

The initial run additionally included an image size of 448 × 448 and a ResNet50; however, this configuration is excluded from subsequent trials due to consistently poor performance observed during the initial run (see Appendix, Figure 6.1).

The confusion matrix (see Figure 4.5) shows that not all ID card of the classes "ID-Card-NOE", "ID-Steiermark", "ID-Card-Tirol_Sanitaeter" und "Unknown" are detected correctly. The class "ID-Card-NOE" exhibits a particularly poor performance.

4.2.3 Metric Learning Approach with Triplet Margin Loss

To improve upon the results achieved by classification methods with CEL, a deep metric learning approach [KB19] is employed. The "Triplet Margin Loss" is chosen as the loss function, and to assign output embeddings to specific classes, a k-NN classifier with $k = 1$ is utilized. Based on promising results in multi-class classification tasks, ResNet18 and ResNet34 architectures are used with input image sizes of 64×64, 112×112 or 224×224 pixels.

[1] Mean Average Precision at an IoU of 0.5

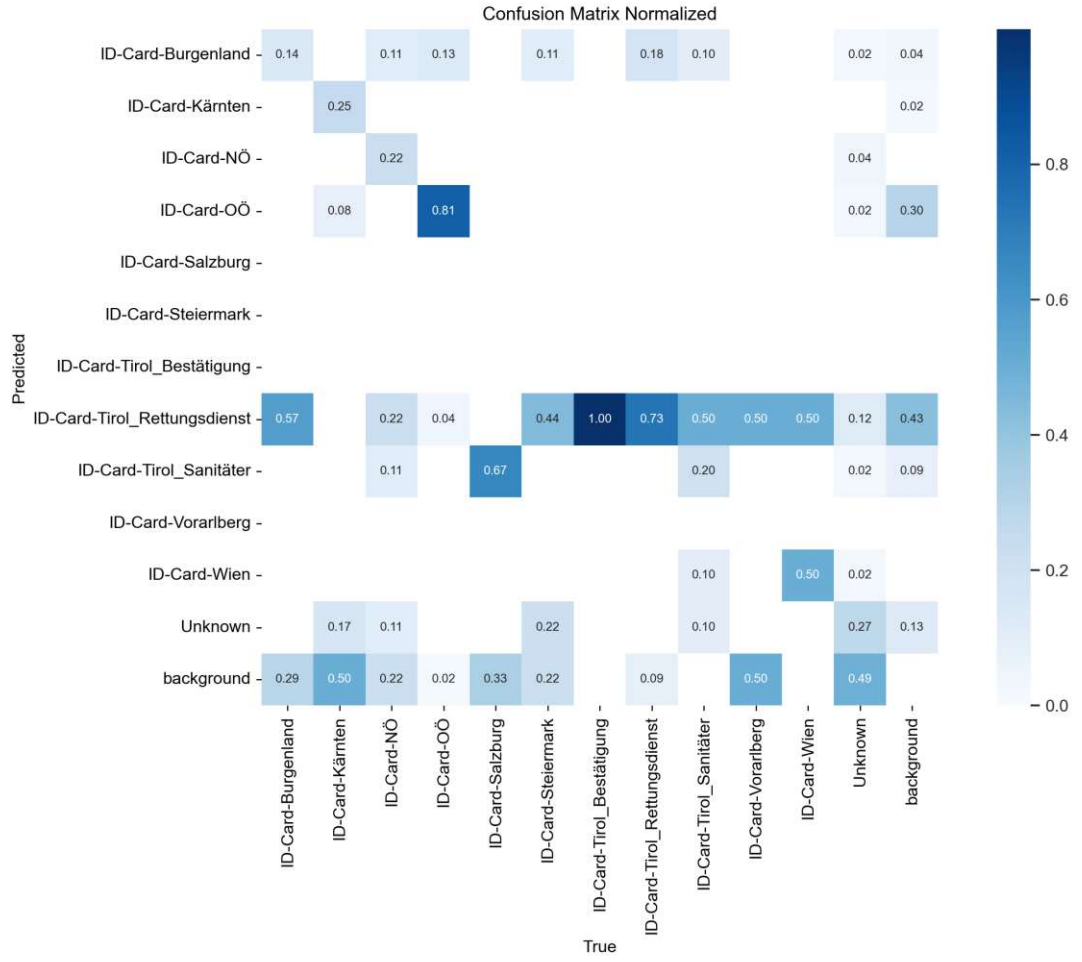


Figure 4.4: The normalized confusion matrix compares the predicted categories of ID/check cards with their actual categories. This indicates that the YOLO model is not capable of accurately classifying ID/check cards into their correct categories. However, an improvement in classification accuracy is observed for classes with a larger number of samples (e.g., "ID-Card-OOE" and "ID-Card-Tirol_Rettungsdienst"). This suggests that a larger sample size may enhance the model's classification performance. However, classes with more samples seem to have more false positives.

The metric learning model achieves a best accuracy of 93.72 % and a best normalized accuracy of 83.25 % (see Table 4.2), falling short of expectations. This approach struggles in particular to identify essential patterns within classes that had fewer training images. Consequently, it is excluded from consideration as the optimal model for ID classification.

Model	64 × 64	112 × 112	224 × 224
Resnet18	**95.17**	93.72	95.17
	93.25	92.08	88.08
Resnet34	95.17	94.20	88.89
	86.25	85.42	80.25

Table 4.1: Performance comparison of ResNet models using CEL at various image sizes. The first value denotes the accuracy, while the second value represents the normalized accuracy.

Model	64 × 64	112 × 112	224 × 224
Resnet18	85.99	84.05	83.09
	68.33	61.16	58.50
Resnet34	86.96	90.34	**93.72**
	67.25	69.75	**83.25**

Table 4.2: The metric learning approach using the Triplet Margin Loss falls short of expectations, with the highest-performing model attaining a maximum accuracy of 93.72% and a maximum normalized accuracy of 83.25%.

4.2.4 Angular Margin-Based Metric Learning Approach

When analyzing the results of the Angular Margin-Based Metric Learning Model (see Table 4.3) in comparison to those of the CEL classification model (see Table 4.1), the Angular Margin-Based model consistently exhibits superior performance across all combinations of model configurations and image sizes. Notably, the Angular Margin-Based approach improves the highest accuracy by 2.9 percentage points relative to the CEL approach, achieving a peak accuracy of 98.07 %. Furthermore, it enhances the normalized accuracy by 3.83 percentage points, attaining a maximum normalized accuracy of 97.08 %.

By looking at the normalized confusion matrix (see Figure 4.6), it becomes evident that also the approach with the built-in ArcFace layer has troubles correctly classifying ID cards of class "ID-Card-NOE".

4.2.5 Classification Result Interpretation

While the suitability of individual training approaches is discussed earlier in this section, no information has been established regarding the potential reasons why variations in approaches, model sizes, and image sizes lead to differences in performance. Furthermore, it remains unclear which optical features serve as the determining factors for these outcomes. Therefore, this subsection examines the potential factors influencing the performance of various model and image sizes, as well as identifies the optical features

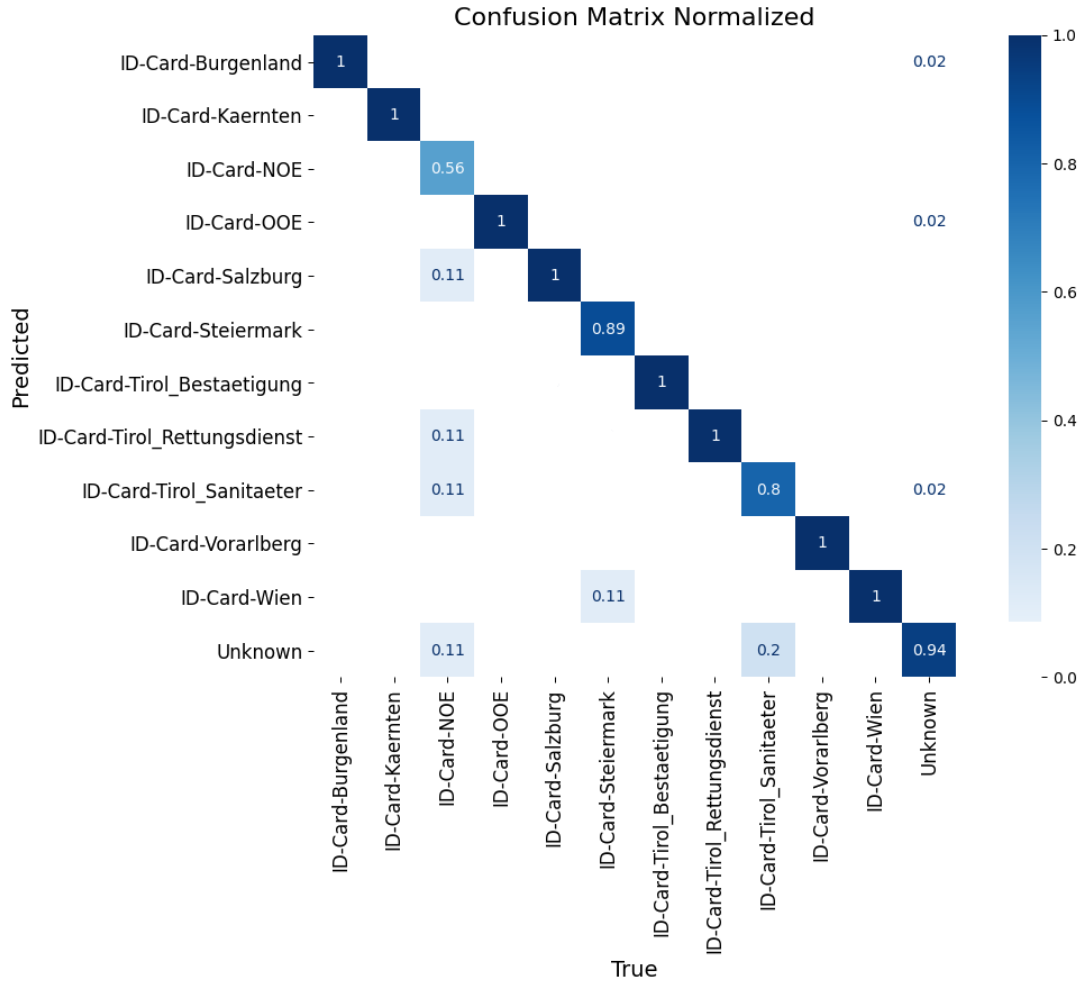


Figure 4.5: The normalized confusion matrix of the model, which achieved an accuracy of 95.17% and a normalized accuracy of 93.25%, compares the predicted classes with the actual belonging classes. It can be observed that the model encounters significant difficulties in correctly classifying ID cards belonging to the "ID-Card-NOE" class.

that are most relevant for classification.

Ablation study

As demonstrated in Section 4.2, the highest accuracy is obtained using the ResNet18 model with an input resolution of 64×64 pixels. At this resolution, the images appear highly pixelated (see Figure 4.7). However, this pixelation may actually enhance the model's accuracy, particularly given the limited number of training samples for some classes (8 out of the 11 classes only have one training image). The reduced resolution introduces a higher level of abstraction, which may aid the model in focusing on visually

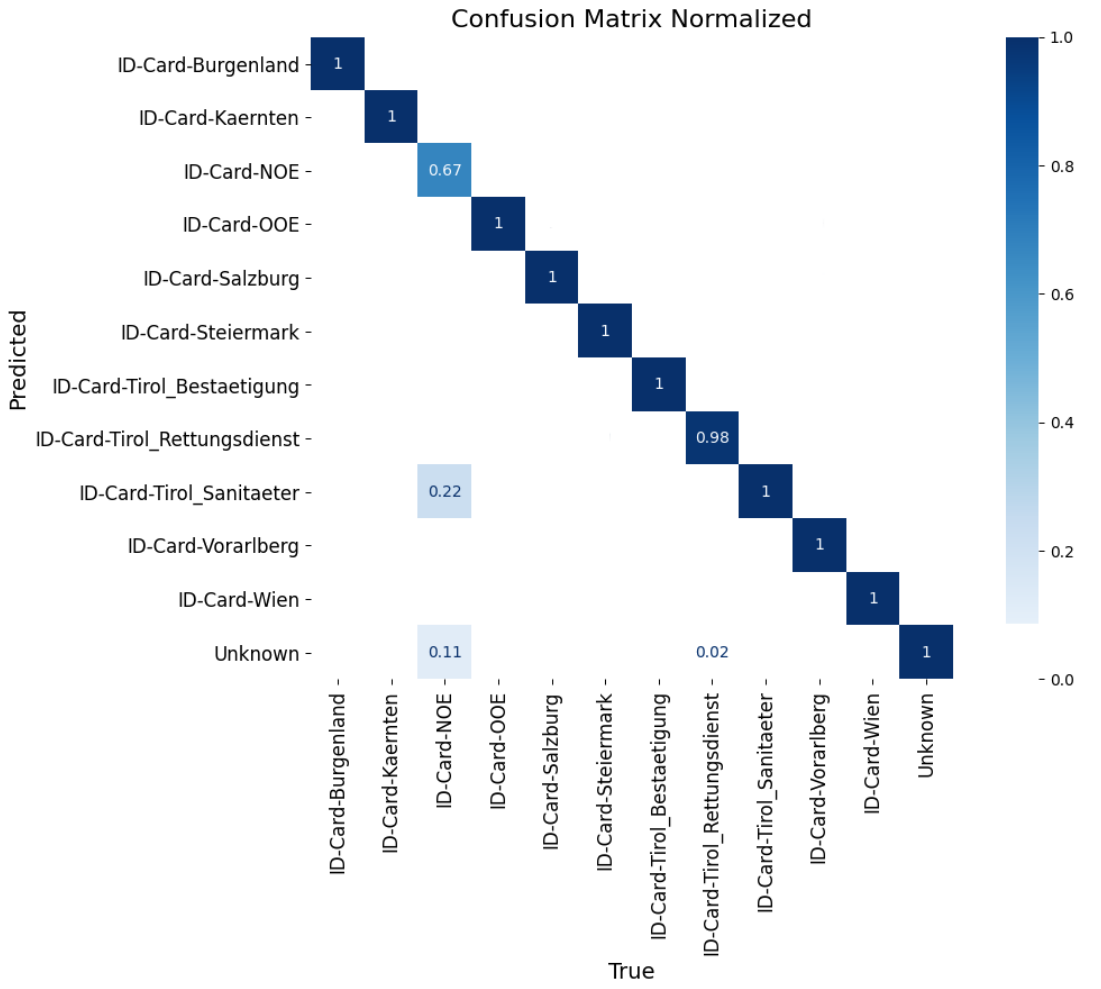
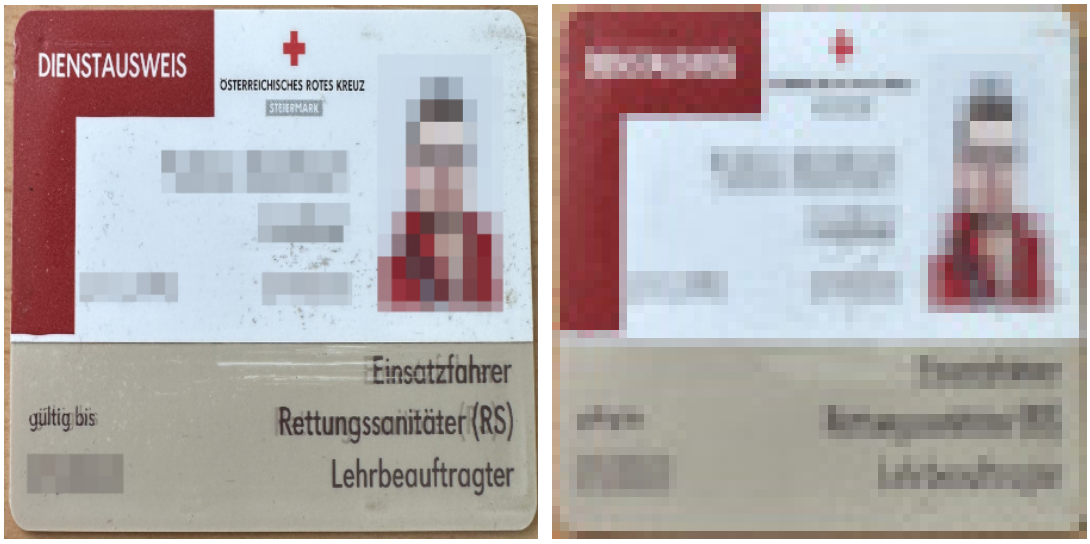


Figure 4.6: The normalized confusion matrix of the model achieving the best performance (accuracy = 98.07% and normalized accuracy = 97.08%) demonstrates that, similar to the CEL model, the primary challenge for the model incorporating the ArcFace layer lies with ID cards belonging to the "ID-Card-NOE" class.

prominent features of the image – such as distinct colored strips – rather than less relevant details, such as the individual's photograph, text specific to the cardholder, background elements, impurities, or signs of wear. This hypothesis is further supported by the observation that, for the 2-class classification task, accuracy improves with both larger model and image sizes, likely due to the greater number of available training samples (27 in total).

Model	64 × 64	112 × 112	224 × 224
Resnet18	**98.07**	95.65	95.65
	97.08	92.42	90.08
Resnet34	96.62	98.07	94.69
	83.75	92.00	92.92

Table 4.3: The results of the models incorporating the ArcFace layer demonstrate a significant improvement compared to the CEL classification approach. Specifically, in comparison to the results of the CEL approach (see Table 4.1), the ArcFace layer increased the highest accuracy by 2.9 percentage points, achieving 98.07%, and enhanced the best normalized accuracy by 3.83 percentage points, reaching 97.08% (indicated by the text highlighted in green). Additionally, improvements were consistently observed across all combinations of model architectures and image sizes.



(a) Image size of 448 × 448 pixels (b) Image size of 64 × 64 pixels

Figure 4.7: Comparison of ID cards with resolutions of 64 × 64 and 448 × 448 pixels: Negative factors for generalization, such as individual text, the person's image, impurities, and signs of wear, are less prominent in the lower-resolution image (64 × 64), thereby reducing the risk of overfitting.

Analyzing Model Behavior Through Visualization

To bring transparency to the underlying mechanisms of the model's inference, the visualization technique *Grad-CAM* [SCD+17] identifies the regions in an image that contribute most significantly to a model's prediction. This is achieved by overlaying a heatmap on the image to highlight the respective areas. Accordingly, in this Section, the Grad-CAM technique is applied to the models trained in the previous Sections to gain

deeper insights into their functioning.

Applying Grad-CAM to the models for ID card classification produces the anticipated outcomes for specific classes. As shown in Figure 4.8, the model accurately identifies regions corresponding to the classes "Burgenland," "Kaernten," and "OOE," thereby enabling these classes to be distinctly recognized from others. However, this analysis also reveals that not all distinctive regions contribute equally to the classification outcome. The model primarily focuses on certain areas of the ID cards while disregarding other regions that also contain unique features relevant to the classification task. This phenomenon is particularly evident in Figures 4.8b and 4.8c, where the card designs are almost entirely unique but are not fully exploited in the model's decision-making process.

(a) The gray circle with the eagle and the red cross represents a distinctive feature of the class "Burgenland".

(b) The ID card of the class "Kaernten" is particularly unique: the model primarily focuses on the tip of the red cross, which is formed by human figures.

(c) The plain white strip with a red cross, bordered by a red block and a red strip, enables the model to distinctly identify the class "OOE".

Figure 4.8: The classes "Burgenland", "Kaernten", and "OOE" are clearly identifiable due to the presence of at least one unique feature or region.

While for the classes "Burgenland," "Kaernten," and "OOE" the decisive areas are very different to all other types in this particular region, for the other classes, the differences in the areas, which are decisive, are more subtle and not as easily comprehensible. Analyzing the decisive areas of the class "Tirol_Sanitaeter" (see Figure 4.9) illustrates the nuanced differences that the model must be capable of recognizing. Although the bottom-right corner of this class appears visually similar to the classes "NOE", "Salzburg" and "Wien", these areas are not identical, enabling the model to differentiate between them.

Examining the results of ID cards of class "Salzburg" highlights the challenges posed by limited training samples (in this case, only one). Under such conditions, the model is at risk to learn irrelevant features. As depicted in Figure 4.10, the model undesirably includes the presence of a face in the upper right corner with its predictions. For the ID card containing an image of a face, the model predominantly focuses on the face and the dark hair to inform its predictions. In contrast, for an ID card featuring a portrait

Figure 4.9: The model successfully identifies the subtle differences between the classes. Notably, the bottom-right corner of the class "Tirol_Sanitaeter" shares significant similarity with other classes; however, distinct subtle differences are present.

image, the model shifts its focus to other regions of the card for making predictions.

(a) Single training ID card for class "Salzburg"

(b) ID card containing only the cardholder's face

(c) ID card containing a portrait of the cardholder

Figure 4.10: Due to the availability of only one training image for the class "Salzburg", the model undesirably interprets the presence of a face in the designated area as a decisive factor when classifying ID cards containing only the cardholder's face. For ID cards containing a portrait image, the model leverages different areas of the ID card for classification.

For class "Tirol_Rettungsdienst", it consistently (unlike other classes where the decisive region within the ID card changes) takes the bottom right corner for decision making (see Figure 4.11). The area it detects contains parts of the image, a strip in different colors (also white, which makes it not distinguishable to the rest of the ID card) and some text.

Except for the lettering "ÖSTERREICHISCHES ROTES KREUZ", which, however, at a resolution of 64 × 64 pixels pixelated beyond recognition, there is no recognizable distinct feature visible.

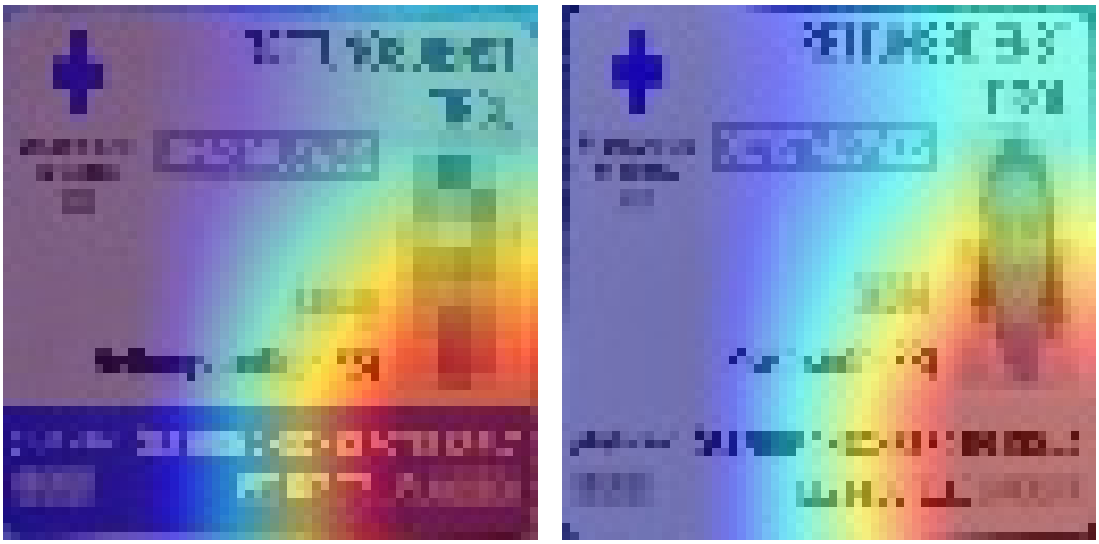


Figure 4.11: ID cards of class "Tirol_Rettungsdienst": The strip at the bottom varies in color depending on the role of the cardholder. However, for the model, the bottom right corner is decisive for classification.

4.2.6 12-Class Approach vs 2-Class Approach

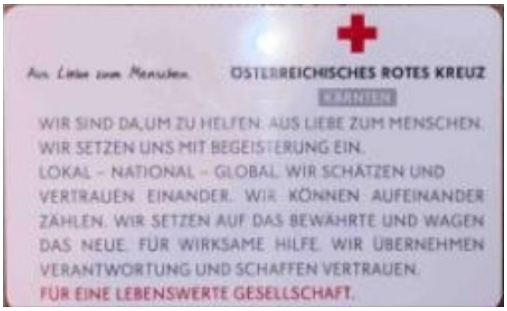
When training the model with one valid and one invalid class only, as presented in Table 4.4, the best-performing model achieves a peak accuracy of 99.03% and a peak normalized accuracy of 98.50%, thereby surpassing the performance of the 12-class approach in terms of accuracy. By absolute numbers, the model misclassified two images: one valid and one invalid sample (see Figure 4.12). An examination of these misclassified samples indicates that, as anticipated, the model trained on all valid ID cards relies on the logo as the decisive region for classification. Due to the high accuracy obtained and the method's reliance on the Red Cross logo, the two misclassified images are highly likely to remain misclassified in subsequent iterations. Consequently, the models were only trained once.

Despite its high accuracy, this approach shows two disadvantages:

1. **Significant focus on the Red Cross logo**: As explicitly shown in Figure 4.13, the model exhibits a dependency on the Red Cross logo (including the lettering beneath the red cross symbol). This focus results in two significant issues:
   - Any ID or check card featuring this logo is likely to be classified as a valid ID card (as illustrated in Figure 4.12a). Consequently, the model cannot reliably

Model	64 × 64	112 × 112	224 × 224	448 × 448
Resnet18	93.72	96.13	98.07	98.55
	87.50	93.50	96.00	97.50
Resnet34	95.65	94.69	94.69	99.03
	91.00	91.50	91.00	98.50

Table 4.4: Results of the models trained using the 2-class approach: Models with larger image sizes demonstrate superior performance. Higher image resolutions offer more detailed features for the network to learn, and the increased availability of training samples enables the network to effectively capture these finer details. This results in improved accuracy.



(a) False Positive check card: The Red Cross logo in the upper right corner is the cause of the erroneous positive classification.

(b) False Negative ID card: The worn off Red Cross logo in the upper left corner is highly likely to be the reason for the misclassification.

Figure 4.12: The two ID/check cards the were misclassified by models trained on the 2-class approach.

distinguish valid ID cards of the Austrian Red Cross from those of other Red Cross organizations.

- Due to the model's heavy reliance on the Red Cross logo, worn or damaged ID cards where the logo is no longer visible are highly likely to be classified as invalid (as illustrated in Figure 4.12b).

2. **Variation in Textual Information Displayed**: Due to the variability in the textual information presented on ID cards, the classification module must identify the relevant class of the ID card in order for the text extraction module to accurately associate the extracted information with the corresponding metadata. To address this challenge, it is necessary to adjust the separation of concerns among the different technologies by integrating OCR capabilities to determine the ID card's classification by extracting the state name directly from the card. However, this approach introduces a potential drawback: if the state name is unreadable for any

reason, the system would fail to classify the ID card correctly, thereby creating a single point of failure.

Figure 4.13: The heatmap generated by Grad-CAM confirms that the logo, including the lettering, constitutes the decisive region for classification.

4.2.7 Assessing Robustness of the Approach through Class Omission

The results presented in Section 4.2.4 demonstrate that the angular margin-based approach effectively classifies nearly all ID/check cards. This confirms its capability to dismiss ID cards that share low degree of similarity. This section examines the robustness of the approach to dismiss ID cards with higher similarity by exposing it to the ID cards of classes omitted from training as described in Section 3.3.5.

An analysis of the results obtained by consecutively omitting each class from the training process reveals that all models achieve consistently high accuracy in their respective best-performing run (see Table 4.5), which is a necessary precondition for a meaningful evaluation of the classification of excluded classes. However, models trained without ID cards from the class "Wien" or, more significantly, from the class "Steiermark" exhibit poor normalized accuracy scores. This issue arises because, in these cases, the best-performing model failed to correctly classify any ID cards for one small class.

For the model excluding the class "Steiermark", neither of the two validation ID cards from the class "Wien" was correctly classified. Similarly, for the model excluding the class "Wien", the single validation ID card from the class "Tirol_Bestaetigung" was not detected. However, since normalized accuracy provides an indication of the model's ability to generalize a class based on its relevant optical characteristics, a low normalized accuracy would only raise concern if the excluded class and the misclassified class exhibited significant visual similarity. In such a scenario, it would be important to investigate whether the model confuses ID cards from the excluded class with those from the visually similar class. In this case, the ID cards from the classes "Steiermark" and "Wien", as well as those from the classes "Wien" and "Tirol_Bestaetigung", do not share high optical resemblance. Therefore, the reduced normalized accuracy does not impact the assessment of the model's robustness.

Excluded Class	Best Accuracy	Normalized Accuracy
Burgenland	96.00	90.36
Kaernten	98.46	99.45
NOE	97.98	97.18
OOE	98.04	94.73
Salzburg	97.55	98.27
Steiermark	96.46	85.91
Tirol_Bestaetigung	98.06	93.27
Tirol_Rettungsdienst	98.15	98.73
Tirol_Sanitaeter	97.46	96.36
Vorarlberg	97.04	95.27
Wien	98.05	89.64

Table 4.5: The table presents the highest accuracy achieved by a model when excluding a specific class. The "normalized accuracy" refers to the normalized accuracy of the model with the highest accuracy. Analysis of the results reveals that the best-performing models across all categories consistently attain high accuracy scores, with the normalized accuracy values also remaining predominantly high.

Classification Analysis of Excluded Classes

When assessing the performance of models on ID cards from classes to which they are not previously trained on (see Table 4.6), the models exhibit limited capability in accurately discerning the relevant optical features enough to reliably rejecting ID cards that share partial or substantial similarities with valid IDs. Quantitatively, the models attain a normalized accuracy of 72.27%, with the "Unknown" class serving as the target label for the tested ID cards. These findings highlight the need for additional measures to enhance the robustness of the approach and prevent IDs with similar optical characteristics from being misclassified as valid classes.

An examination of the individual results highlights a notable affinity between the classes "Salzburg" and "NOE" caused by the significant similarities in the optical features of their ID cards. Specifically, the bottom third of the two cards is nearly identical, with the only distinction being the text in the gray box, which reads "SALZBURG" or "NIEDERÖSTERREICH" respectively. Furthermore, the lettering "DIENSTAUSWEIS" ("Company ID Card" in German) appears identically in the top-left corner of both cards, and the term "Dienstnr.:" ("Service Number" in German) is located in a similar position and is optically identical. Additionally, the placement of the individual's photograph is consistent between the two classes. Given these shared characteristics, this misclassification is expected, particularly when the model has not been explicitly trained to distinguish between the two.

However, misclassifications are not limited to classes with strong visual resemblance. For example, 50% of the ID cards from class "Burgenland" are erroneously assigned to

the class "Tirol_Rettungsdienst". This misclassification occurs despite the two classes sharing minimal resemblance, aside from a broad, single-colored strip at the bottom of the card.[2].

Another observation is that 17% of the ID cards from the "OOE" class are misclassified as belonging to the "Tirol_Sanitaeter" class, even though these two categories exhibit very little similarity. The only common feature between the two is the text "ÖSTERREICHISCHES ROTES KREUZ" ("Austrian Red Cross" in German) printed in two lines beneath the logo, even though the specific locations referenced on the cards differ significantly. However, the two-line text layout is also a characteristic of ID cards from the class "Kaernten". A plausible explanation for why the model assigns the classes to "Tirol_Sanitaeter" instead of "Kaernten" could be the higher number of available samples for the class "Tirol_Sanitaeter" in the dataset.

Excluded Class	**Burgenland**	**Kaernten**	**NOE**	**OOE**	**Salzburg**	**Steiermark**	**Tirol_Bestaetigung**	**Tirol_Rettungsdienst**	**Tirol_Sanitaeter**	**Vorarlberg**	**Wien**	**Unknown**
Burgenland	–							0.50				0.50
Kaernten		–										1.00
NOE			–		1.00							
OOE				–				0.17				0.83
Salzburg			0.50		–				0.50			
Steiermark						–						1.00
Tirol_Bestaetigung							–					1.00
Tirol_Rettungsdienst						0.02		–	0.04		0.08	0.87
Tirol_Sanitaeter								0.08	–		0.17	0.75
Vorarlberg										–		1.00
Wien											–	1.00

Table 4.6: The table illustrates the percentage distribution of ID cards from excluded classes across various predicted classes. Each row corresponds to an excluded class, while each column represents a predicted class. The "Unknown" class serves as the target class. A higher percentage of ID cards classified as "Unknown" reflects greater robustness of the classification method when handling ID cards with similar optical characteristics.

[2]The strip for "Tirol_Rettungsdienst," varies in color depending on the role, rather than being consistently red as depicted in the example images in Figure 3.1

Incorporating Confidence Measures in Predictions

The highest probability score produced by a model can be interpreted as the model's confidence in its inference. Accordingly, applying a confidence threshold can help filter out ID cards that are incorrectly classified as valid. Table 4.7 presents a breakdown of which percentage of such incorrectly classified valid ID cards, that the model was not trained on, fall within different confidence ranges.

Excluded Class	0–80	80–90	90–99	99.0–99.9	99.90–99.95	99.95–100
Burgenland	1		2	1		
NOE				1	1	8
OOE	10		1			
Salzburg	1		1			2
Tirol_Rettungsdienst	2	1	3	1		
Tirol_Sanitaeter	1	1	1			
Total	15 (38.46%)	2 (5.13%)	8 (20.51%)	3 (7.69%)	1 (2.56%)	10 (25.64%)

Table 4.7: Confidence scores for all ID cards from omitted classes that were erroneously assigned to one of the remaining valid classes. The intervals are defined with the upper bound included and the lower bound excluded. The results also highlight the visual similarity between the classes "NOE" and "Salzburg," as misclassified ID cards from one class to the other exhibit very high confidence scores.

To determine this threshold in a systematic and reasonable manner, it is necessary not only to analyze the confidence scores of True Negatives but also to evaluate the confidence levels of True Positive ID cards. Specifically, this involves examining the confidence scores of all true positive predictions across all classes (excluding the "Unknown" class, as ID cards below the threshold are classified as invalid and thus categorized under the "Unknown" class). The confidence levels (see Table 4.8) are derived from the outputs of the best-performing model, as identified in Section 4.2.4. The dataset utilized for the confidence assessment includes approximately 15% of the ID cards that were previously used for training. However, this does not introduce bias into the results, as ID cards with very high confidence scores have minimal or no impact on evaluating the ratio of True Negatives to False Negatives.

After calculating the confidence scores for ID cards classified as valid – both correctly (i.e., by the best-performing model identified in Section 4.2.4) and erroneously (i.e., misclassified as valid by models excluding specific classes) – various thresholds can be assessed by evaluating the proportion of ID cards that would be correctly or incorrectly rejected when these thresholds are used as validity criteria. By analyzing the ratio of true negatives to false negatives across different thresholds, as shown in Table 4.9, a minimum confidence threshold of 80% achieves the highest ratio of 7.5. Therefore, an 80% threshold represents a reasonable trade-off.

ID Class	0–80	80–90	90–99	99.0–99.9	99.90–99.95	99.95–100
Burgenland	1	1	1	4	1	
Kaernten						13
NOE			1	1		5
OOE						63
Salzburg				1		3
Steiermark						10
Tirol_Bestaetigung					1	1
Tirol_Rettungsdienst						52
Tirol_Sanitaeter						12
Vorarlberg	1			1	3	
Wien		1	1			1
Total	2 (1.12%)	2 (1.12%)	3 (1.68%)	7 (3.91%)	5 (2.79%)	160 (89.39%)

Table 4.8: Confidence levels of all correctly classified ID cards using the best-performing model identified in Section 4.2.4. The results demonstrate that the majority of correctly classified ID cards are associated with very high confidence scores.

	80%	**90%**	**99%**	**99.9%**	**99.95%**
Ratio (TN/FN)	7.5 (15 / 2)	4.25 (17/4)	3.57 (25 / 7)	2 (28 / 14)	1.53 (29 / 19)

Table 4.9: Ratio of True Negatives (TN) to True Positives (TP) when discarding ID cards exhibiting a confidence below a certain threshold: A threshold of 80% confidence achieves the highest ratio of 7.5. Specifically, 15 out of 17 ID cards with confidence scores below 80% are correctly discarded (True Negatives), while 2 are incorrectly discarded (False Negatives). These results suggest that a threshold of 80% represents a reasonable tradeoff.

4.3 Text Extraction via OCR

The final task of the pipeline is the extraction of the ID card's textual data. Primarily, this thesis compares the performance of three OCR technologies: Tesseract, EasyOCR and PaddleOCR. Table 4.10 demonstrates that while Tesseract is notably fast, its accuracy is insufficient, rendering it unsuitable for the task of text extraction from ID cards. Although EasyOCR achieves a considerable improvement in accuracy compared to Tesseract, it is significantly surpassed by PaddleOCR in both speed and accuracy.

4.3.1 Combining Models for enhanced Perfomance

Although PaddleOCR demonstrates superior performance over EasyOCR in absolute metrics, it does not provide explicit information about specific fields. Consequently, there is a possibility that EasyOCR may detect certain fields that PaddleOCR misses. Therefore, employing EasyOCR as a complementary fallback mechanism when PaddleOCR fails

OCR Technology	Average Speed	Correct Fields	Correct ID Cards
Tesseract	0.45s	36.47%	24.04%
EasyOCR	7.28s	80.71%	57.38%
PaddleOCR	1.86s	93.35%	82.51%

Table 4.10: Comparison of OCR technologies in terms of speed and accuracy: Speed is quantified as the execution time (in seconds) per image, while accuracy is assessed based on two metrics: the ratio of relevant fields correctly identified and the proportion of ID cards where all relevant fields were accurately captured. Fields for which the text could not be clearly determined or required domain-specific knowledge for inference (e.g., ambiguity between characters such as 0, 8, or 9, which may be resolvable only when interpreted as part of a date) were excluded from the evaluation.

to correctly detect a field is examined. To evaluate the combined performance of both models, the process begins with running the PaddleOCR model. If not all fields are accurately extracted, the EasyOCR model is subsequently employed to capture the remaining fields, which are then matched with the text detected by EasyOCR. This combined approach improves the ratio of correctly extracted fields to 95.30% and the ratio of correctly identified ID cards to 87.43%.

4.3.2 Reasons for Erroneous Recognition

An analysis of the misclassified images reveals several root causes for recognition errors:

1. **Pixelation**: Insufficient image resolution issue is particularly problematic when interpreting dates, as numerical characters are more prone to misclassification compared to letters. Additionally, the dots separating day, month, and year often become indistinct or nearly invisible in highly pixelated images, making correct recognition almost impossible.

2. **Signs of Wear**: While a human observer might still be able to infer the intended meaning of the text, for OCR models focused solely on character detection, this task becomes exceedingly challenging (see Figure 4.14).

3. **Umlauts**: Despite the models being trained on German text, they struggle to accurately recognize umlauts.

4. **Additional Factors**: Additional factors contributing to insufficient recognition of human-readable text include partially covered text due to fingertips or ID cases, light reflections, or overlapping print layers, such as updated printings over fragments of initial printings that remain visible.

(a) The central segment of the letter "m" is faintly visible; however, the OCR model fails to recognize it.

(b) Given that the displayed character is a printed digit, the exclusion of other possibilities based on the visible portion indicates that it is the digit "6".

(c) While the digit "7" is correctly recognized, the digit "3" is not.

Figure 4.14: Although the intention of the displayed letters and digits can inferred by a human spectator, the model is not capable of correctly recognizing it.

4.3.3 Improving Accuracy with Contextual Understanding

Due to MLLM's capability to process images and text, this Section examines the text extraction capabilities of "Llama 3.2 Vision". For ID cards where text is not correctly extracted using PaddleOCR, Llama successfully identified **90.10% of the individual fields**, and for **65.22%** of the ID cards, **all fields were recognized correctly**. When the fields accurately identified by Llama were combined with those recognized by PaddleOCR, the overall text extraction performance improved, enabling 96.72% of all ID cards to be correctly processed. This improvement demonstrates a notable enhancement in recognition capability, attributable to Llama's contextual understanding. This advantage is particularly evident in its ability to reconstruct dates in the correct format (e.g., with separating dots) even when the separators are entirely absent due to pixelation or inferences of digits that are mostly obscured, making their identification impossible based solely on the visible portions of their shapes.

However, this approach has a notable limitation: the computational cost of Llama is substantially higher than that of PaddleOCR. Specifically, the average runtime of the Llama model on a single image is measured at 257.62 seconds. As a result, deploying a computationally expensive MLLM in real-time applications is only feasible if adequate hardware resources are available or, in cases where privacy concerns are not a priority, the MLLM can be executed on external hardware (e.g., via an API).

4.3.4 Error Discrimination

A comparison of the text detected by PaddleOCR and the ground truth indicates that the majority of error manifestations can be categorized into the following types:

1. Missing separating dots in dates (e.g., "01.012020" instead of "01.01.2020").
2. Missing separating slash in dates (e.g., "012020" instead of "01/2020").
3. Incorrect recognition of umlauts.

4. One letter not or incorrectly recognized (i.e., Levenshtein distance of 1).

5. One digit not or incorrectly recognized (i.e., Levenshtein distance of 1).

Table 4.11 illustrates that only 24.39% of the incorrectly read fields fall outside the identified patterns. Consequently, it is feasible to establish rules that tolerate a certain error margin. When examining the individual error patterns in detail for practical applicability, errors categorized under points 1–4 can generally be regarded as minor deviations from the ground truth. However, point 2 can only be addressed by processing each 6-digit string and inserting a slash at the appropriate position. This introduces a minor risk of misinterpreting a numerical value as a date. A similar issue, though mitigated, arises from point 1.

For point 5, the assessment depends on the specific case. In scenarios involving service numbers and dates of birth, where reference data is available, minor errors – defined as errors with a Levenshtein distance of 1 – can be addressed similarly to the errors in points 1–4. However, determining the correct validity date from an invalid one requires a case-specific evaluation. It can be reasoned that, based on the positions of separating symbols (e.g., dots or slashes) and the limited range of possible numerical values in dates, an incorrect digit may be identified. The impact of such an error depends on the digit's position within the date.

	Separating Dots	**Separating Slash**	**Umlaut**	**One Letter**	**One Digit**	**Others**
Distribution of errors	12.20 %	7.32 %	9.76 %	29.27 %	17.07 %	24.39 %

Table 4.11: Distribution of all errors across the listed error manifestation categories. Each field within a section exclusively exhibits the specified error pattern; otherwise, it is classified under the category "Others".

The significance of validity date components decreases hierarchically from the year to the month and, where applicable, to the day. Specifically, only if the year corresponds to the current year, the month and day (if provided) are critical. Errors in the year are typically difficult or impossible to reconstruct, whereas errors in the month or day can often be mitigated due to the restricted range of valid values. For such corrections, an incorrect digit can be replaced with the lowest possible valid value to ensure the date remains valid (e.g., transforming "99/2024" to "09/2024"). The resulting date is then assessed for validity.

Applying an error margin tolerance that includes errors from points 1-4 or reconstructable cases from point 5 to the output of PaddleOCR, **97.89% of all fields and 93.99% of all ID cards** are correctly read.

CHAPTER 5

Conclusion

5.1 Summary

This thesis investigates the feasibility of utilizing deep learning techniques to automate the verification of ID cards, with a focus on their optical characteristics. The dataset used for this study consists of 11 distinct types of valid ID cards [1] and one class representing invalid ID cards or other check cards. Due to their ability to analyze patterns in the images [LBH15], in this study, CNNs are employed to detect relevant features and the assign of each ID card to its corresponding type. The complexity of this task arises from the uneven distribution of samples across classes and variations in the image quality of the ID cards. This study does not address the authentication of individuals through face comparison with the photograph on the ID card, nor does it assess the model's ability to identify and reject counterfeit ID cards.

For verifying the ID cards, this thesis proposes a pipeline that integrates three steps:

1. Segmentation of the ID/check card from its background
2. Classification of the ID/check card as valid or invalid
3. Extraction of the textual information on the ID card

For segmentation, this study evaluates the capability of the object detection algorithm YOLOv8 to accurately segment ID cards from their backgrounds. To evaluate segmentation quality, the Recall@IoU-MinIS metric is proposed, which incorporates a minimum intersection score in conjunction with Recall@IoU to emphasize the preservation of the complete ID card without cropping significant portions of the ID card. The analysis

[1] Here, "valid" refers to ID cards issued by the Austrian Red Cross.

demonstrates that a Recall at an IoU threshold of 0.8 and a minimum intersection score of 0.9 provides a reliable measure of segmentation accuracy. The findings indicate that the YOLOv8 model successfully segments 90% of the ID cards with precision. Furthermore, the study identifies key patterns in segmentation errors, such as insufficient contrast between the ID card and the background, as well as the unintentional cropping of relevant regions due to sharp edges in the ID card design.

The classification module represents a cornerstone of this thesis. Given the dataset imbalance, its evaluation utilizes normalized accuracy (i.e., the mean of the accuracies across all classes) in addition to the overall accuracy. It investigates several classification approaches: YOLOv8-based classification, Cross-Entropy Loss, metric learning using Triplet Margin Loss, and Angular Margin-Based Metric Learning (ArcFace). The results imply the inadequacy of YOLOv8 as a standalone classification tool due to its limited performance in distinguishing ID card classes. Conversely, the ResNet architectures combined with ArcFace demonstrated superior performance, achieving a peak accuracy of 98.07% and a normalized accuracy of 97.08%. This outcome validates the effectiveness of angular margin-based metric learning for multi-class classification tasks involving nuanced visual features.

In addition to presenting the results of the performance evaluation of the classification model, this study conducts an ablation study on image and model sizes, examines the decisive factors influencing the model's predictions using the Grad-CAM visualization technique, contrasts a 2-class (i.e., only valid and invalid class) classification approach to the approach with a 12-class classification (i.e., where each ID card type is treated as a separate class), and evaluates the model's robustness when applied to ID cards with high visual similarity:

1. **Ablation study**: For 12-class approach, a lower image resolution (64×64) combined with a reduced model size (ResNet18) demonstrates the best performance. Conversely, for binary classification tasks, higher image resolutions (448×448) and larger model sizes (ResNet34) result in improved accuracy. In this scenario, the use of reduced image resolution may help compensate for the limited number of training images in certain classes by providing a higher level of abstraction.

2. **Model Behavior Visualization**: The visualization effectively demonstrates that the model relies on distinctive features for classification in certain classes. However, it also reveals that for classes with limited training samples, the model tends to use regions specific to the individual cardholder (e.g., the individual's image) for classification.

3. **12-Class vs 2-Class Approach**: The evaluation of the 2-class approach demonstrates superior performance compared to the 12-class approach, achieving a peak accuracy of 99.03% and a peak normalized accuracy of 98.50%. Additionally, the study addresses the limitations associated with this approach, both in terms of reliance on the Red Cross logo and the proposed pipeline.

4. **Robustness of the approach**: The thesis demonstrates that the model exhibits limited robustness when tested with visually similar ID cards (simulated by excluding a class and exposing its instances to the model). The model achieves a normalized accuracy[2] of 72.27%. Furthermore, the thesis discusses how incorporating a minimum confidence score enhances the model's robustness, with a threshold of 80% identified as yielding the optimal balance between true negatives and false negatives when rejecting ID cards classified with confidence below this value.

For the extraction of textual information from ID cards, three OCR techniques are evaluated: Tesseract, EasyOCR, and PaddleOCR. Among these, PaddleOCR demonstrates the highest accuracy, correctly identifying 93.35% of the relevant fields and achieving full correctness for 82.51% of ID cards. EasyOCR achieves a lower accuracy, correctly extracting 80.71% of all fields. However, combining the outputs of PaddleOCR and EasyOCR significantly improves accuracy, with 95.30% of all fields correctly extracted and 87.43% of ID cards having all fields extracted correctly.

The study identifies key factors contributing to recognition errors, including pixelation, signs of wear, the presence of umlauts, and additional factors. To further enhance performance, the contextual understanding capabilities of the "Llama 3.2 Vision" model are utilized. When tasked with processing only those ID cards that PaddleOCR failes to read accurately, the model correctly extracts 90.10% of all fields. By integrating Llama with PaddleOCR, the combined system achieves an overall accuracy of 96.72% for fully reading ID cards. However, the computational cost of using Llama exceeds that of PaddleOCR by approximately 138 times.

Finally, the study demonstrates that applying an error margin tolerance improves the performance of PaddleOCR, achieving a correctness rate of 97.89% for all fields and 93.99% for ID cards with all fields correctly extracted.

5.2 Future Work

This thesis has demonstrated the feasibility of utilizing deep learning methodologies for ID card verification, with a focus on optical characteristics. However, several areas for future research and development remain:

Enhance Segmentation In this pipeline, the YOLOv8 model is trained to perform object detection using bounding boxes. However, because these bounding boxes are axis-aligned with the edges of the image, ID cards that are skewed within the image result in cropped outputs that remain skewed. Consequently, the cropped ID card, which serves as input for classification, either includes portions of the background or has sections of the ID card truncated. To address this issue, it is worth investigating the effectiveness

[2]Since this study evaluates the similarity between classes, normalized accuracy serves as a more meaningful metric than overall accuracy.

of semantic segmentation using polygons instead of bounding boxes. Alternatively, the performance of other segmentation techniques, such as the Segment Anything Model (SAM) [KMR+23], can be explored.

Fraud Detection As stated, this thesis does not address the detection of forged ID cards. Consequently, the pipeline does not incorporate any countermeasures against such forgeries. Therefore, in this context, applying the pipeline to a counterfeit ID card that closely resembles a legitimate ID card will result in its classification as valid. In future work, the feasibility of integrating fraud detection into the verification pipeline could be explored. This would involve extending the dataset to include forged ID cards, such as those generated by AI from an authentic ID card, manipulated images of ID cards (e.g., altered text), or photocopies of ID cards.

Employ Vision Transformers for Classification Liu et al. [LSB+21] propose an approach to improve the performance of Vision Transformers when working with limited training data. Given the ability of Vision Transformers to capture global relationships, this technology could be explored to determine whether it enhances robustness in distinguishing ID cards that differ minimally from those the model has been trained to recognize as valid.

MLLMs for Classification Future work could explore using MLLMs in a Retrieval-Augmented Generation system for classification. By storing ID card embeddings in a vector database, this approach could eliminate the need for model training while simultaneously enabling text extraction.

CHAPTER 6

Appendix

6.1 Results Cross-Entropy Loss

The following tables present the results of a ResNet model employing Cross-Entropy Loss when training on 12 inference classes.

Model	64 × 64	112 × 112	224 × 224	448 × 448
Resnet18	92.27	93.72	95.17	91.30
	80.83	92.08	88.08	79.75
Resnet34	91.30	91.30	88.89	84.54
	79.67	76.83	80.25	67.33
Resnet50	89.37	60.39	83.57	76.81
	79.97	26.83	79.75	40.08

Table 6.1: First Run (Cross-Entropy Loss)

Model	64 × 64	112 × 112	224 × 224
Resnet18	95.17	92.75	94.20
	87.83	86.91	92.00
Resnet34	92.75	83.57	87.92
	85.42	66.92	80.92

Table 6.2: Second Run (Cross-Entropy Loss)

Model	64 × 64	112 × 112	224 × 224
Resnet18	95.17	93.72	93.72
	93.25	75.33	86.33
Resnet34	95.17	94.20	85.99
	86.25	85.42	86.25

Table 6.3: Third Run (Cross-Entropy Loss)

6.2 Results Angular Margin-Based Metric Learning

The following tables present the results of a ResNet model employing Angular Margin-Based Metric Learning when training on 12 inference classes.

Model	64 × 64	112 × 112	224 × 224
Resnet18	95.17	95.17	92.75
	89.75	91.42	85.00
Resnet34	96.62	89.85	94.69
	83.75	91.08	92.92

Table 6.4: First run (Angular Margin-Based Metric Learning)

Model	64 × 64	112 × 112	224 × 224
Resnet18	98.07	94.69	95.65
	97.08	87.42	90.08
Resnet34	96.14	94.69	93.72
	81.25	92.83	86.08

Table 6.5: Second run (Angular Margin-Based Metric Learning)

Model	64 × 64	112 × 112	224 × 224
Resnet18	97.10	95.65	95.17
	92.58	92.42	87.42
Resnet34	96.14	98.07	93.24
	94.17	92.00	84.75

Table 6.6: Third run (Angular Margin-Based Metric Learning)

6.3 Results Class Omission

The following tables present the results of a ResNet model employing Angular Margin-Based Metric Learning, with the exclusion of one class.

Excluded Class	Accuracy	Normalized Accuracy
Burgenland	96.50	95.00
Kaernten	94.87	88.73
NOE	95.96	96.36
OOE	92.81	87.55
Salzburg	97.55	98.27
Steiermark	95.45	85.27
Tirol_Bestaetigung	94.66	88.64
Tirol_Rettungsdienst	94.44	87.73
Tirol_Sanitaeter	93.91	92.55
Vorarlberg	95.07	88.55
Wien	96.59	91.27

Table 6.7: First run with class omission

Excluded Class	Accuracy	Normalized Accuracy
Burgenland	96.00	90.36
Kaernten	97.44	92.45
NOE	97.98	81.64
OOE	98.04	94.73
Salzburg	97.05	96.09
Steiermark	94.95	93.27
Tirol_Bestaetigung	94.17	89.90
Tirol_Rettungsdienst	97.53	96.90
Tirol_Sanitaeter	97.46	96.36
Vorarlberg	97.04	91.55
Wien	97.56	93.91

Table 6.8: Second run with class omission

Excluded Class	Accuracy	Normalized Accuracy
Burgenland	94.00	79.18
Kaernten	98.46	99.45
NOE	97.98	97.18
OOE	90.85	81.18
Salzburg	95.10	91.27
Steiermark	96.46	85.91
Tirol_Bestaetigung	98.06	93.27
Tirol_Rettungsdienst	98.15	98.73
Tirol_Sanitaeter	96.45	85.73
Vorarlberg	97.04	95.27
Wien	98.05	89.64

Table 6.9: Third run with class omission

Overview of Generative AI Tools Used

Übersicht verwendeter Hilfsmittel

List of Figures

List of Tables

Bibliography

[CMS+20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[CPIS19] Adulwit Chinapas, Pattarawit Polpinit, Narong Intiruk, and Kanda Runapongsa Saikaew. Personal verification system using id card and face photo. *International Journal of Machine Learning and Computing*, 9(4):407–412, 2019.

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[GSRK21] Manish Kumar Gupta, Ronak Shah, Jitesh Rathod, and Ajai Kumar. Smartidocr: Automatic detection and recognition of identity card number using deep networks. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 267–272, 2021.

[JCQ23] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023.

[KB19] Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

[Kin14] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KMR+23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[LLH+24] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[LSB+21] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021.

[LVS+21] Rodrigo Lara, Andres Valenzuela, Daniel Schulz, Juan Tapia, and Christoph Busch. Towards an efficient semantic segmentation method of id cards for verification systems. *arXiv preprint arXiv:2111.12764*, 2021.

[RDGF16] 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 (CVPR)*, June 2016.

[RSHS24] P Praneeth Reddy, P Sai Shruthi, P Himanshu, and Tripty Singh. License plate detection using yolo v8 and performance evaluation of easyocr, paddleocr and tesseract. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2024.

[SCD+17] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Sch15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[Ult23] Ultralytics. Yolo performance metrics guide. `https://docs.ultralytics.com/guides/yolo-performance-metrics/#object-detection-metrics`, 2023. Accessed: 2024-10-31.

[VAT20] Andres Valenzuela, Claudia Arellano, and Juan E Tapia. Towards an efficient segmentation algorithm for near-infrared eyes images. *IEEE Access*, 8:171598–171607, 2020.

[WBL23] 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.

[WCL+24] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024.

[WXW+19] Xing Wu, Jianxing Xu, Jianjia Wang, Yufeng Li, Weimin Li, and Yike Guo. Identity authentication on mobile devices using face verification and id image recognition. *Procedia Computer Science*, 162:932–939, 2019. 7th International Conference on Information Technology and Quantitative Management (ITQM 2019): Information technology and quantitative management based on Artificial Intelligence.