sustainability

MDPI

Article

The Industrial Application of Artificial Intelligence-Based Optical Character Recognition in Modern Manufacturing Innovations

Qing Tang [1], YoungSeok Lee [1] and Hail Jung [2,*]

[1] Data Science Group, INTERX, Ulsan 44542, Republic of Korea; tangqing@interxlab.com (Q.T.); ys.lee@interxlab.com (Y.L.)
[2] Department of Business Administration, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
* Correspondence: hail95@seoultech.ac.kr; Tel.: +82-10-2994-7527

Abstract: This paper presents the development of a comprehensive, on-site industrial Optical Character Recognition (OCR) system tailored for reading text on iron plates. Initially, the system utilizes a text region detection network to identify the text area, enabling camera adjustments along the x and y axes and zoom enhancements for clearer text imagery. Subsequently, the detected text region undergoes line-by-line division through a text segmentation network. Each line is then transformed into rectangular patches for character recognition by the text recognition network, comprising a vision-based text recognition model and a language network. The vision network performs preliminary recognition, followed by refinement through the language model. The OCR results are then converted into digital characters and recorded in the iron plate registration system. This paper's contributions are threefold: (1) the design of a comprehensive, on-site industrial OCR system for autonomous registration of iron plates; (2) the development of a realistic synthetic image generation strategy and a robust data augmentation strategy to address data scarcity; and (3) demonstrated impressive experimental results, indicating potential for on-site industrial applications. The designed autonomous system enhances iron plate registration efficiency and significantly reduces factory time and labor costs.

Keywords: artificial intelligence; optical character recognition; manufacturing innovation; manufacturing industrial; real-world application

Citation: Tang, Q.; Lee, Y.; Jung, H. The Industrial Application of Artificial Intelligence-Based Optical Character Recognition in Modern Manufacturing Innovations. *Sustainability* **2024**, *16*, 2161. https://doi.org/10.3390/su16052161

Academic Editor: Jun (Justin) Li

Received: 13 December 2023
Revised: 26 February 2024
Accepted: 26 February 2024
Published: 5 March 2024

1. Introduction

The advent of Industry 4.0 marks a paradigm shift in manufacturing, integrating cutting-edge digital technologies into established manufacturing practices. This transformation underscores the indispensable role of automation in modern manufacturing workflows as highlighted in seminal works [1–8]. Contemporary advancements in manufacturing predominantly focus on two domains: automation and robotics, aimed at streamlining repetitive tasks, and Artificial Intelligence (AI) coupled with Machine Learning (ML), leveraged to enhance production efficiency and optimize logistical decision-making [9,10]. These innovations are instrumental in elevating production efficiency, reducing time, and curtailing labor expenses because the demand for rapid production cycles and the need to optimize resource allocation in manufacturing have made efficiency a critical concern.

A critical component in this technological evolution is an autonomous product information registration system. Traditional manual or semi-automated registration systems struggle with the dynamic and often harsh industrial environments, leading to delays and increased cycle times. The traditional manual methods are not only labor-intensive but also prone to human error, leading to significant operational inefficiencies. Our research is motivated by the industry's urgent need to automate the registration process, thereby

minimizing human intervention and enhancing the reliability of the process. This works focus on designing an autonomous and accurate on-site iron plates registration system for speed up iron plates identification, registration, and tracking by utilizing Optical Character Recognition (OCR) technology. The proposed OCR system addresses these inefficiencies and significantly reducing factory time and labor costs.

The OCR systems are adept at extracting textual data from images, documents, or physical objects and transcribing optical text into digital formats, thus playing a pivotal role in the modernization of manufacturing processes [11–16]. The implementation of OCR technology spans diverse industrial sectors, including automotive [16], iron [12,15], printed circuit board (PCB) [11,14], and pharmaceuticals [17]. The efficacy of OCR in these sectors can be attributed to three primary reasons:

1. OCR systems significantly enhance the efficiency of production lines by providing continuous operation, which markedly contrasts with the slower, more error-prone manual inspection processes, particularly in high-volume manufacturing environments.
2. These systems offer consistent and accurate text-reading capabilities, ensuring precise product identification, a critical requirement often compromised in manual inspections, especially at scale.
3. The ability of OCR systems to rapidly convert optical character data into digital format is increasingly crucial as manufacturers integrate Internet of Things (IoT) and Big Data analytics into their operations. This integration, facilitated by OCR, enables real-time data management and analysis, a cornerstone of Industry 4.0.

This paper presents a comprehensive system for segmenting and recognizing codes printed on iron plates—critical for tracking vital information such as company brand, manufacturing date, and product specifications in industrial settings. Figure 1 illustrates the onsite photos and schematic of the iron plate registration system. The traditional approach involves workers manually inputting these data after visual inspection, a process fraught with inefficiencies. An industrial-grade OCR system could significantly streamline this process. However, the deployment of OCR in industrial settings is not without its challenges:

1. Data Availability: There is a lack of suitable public data for training the OCR system. Additionally, the collection and labeling of on-site data are challenges regarding both time and financial resources.
2. The on-site data quality is poor as illustrated in Figure 2. It causes difficulty in both training and testing. Specific issues include the following:
    - Figure 2a: High background noise.
    - Figure 2b: Variable text sizing due to the varying distances between the camera and the target text.
    - Figure 2c: Issues with illumination, leading to reflections or shadows.
    - Figure 2d: Wear and tear: Text on iron plates erodes due to water stains and material wear in outdoor and factory settings. This degradation is exacerbated when the plates are stacked, leading to further deterioration.
    - Figure 2e: The use of varied, uncommon, and non-standard fonts renders public font datasets and pre-trained models unsuitable for direct application in our specific task.
    - Figure 2f: Non-frontal shooting angles result in severe perspective transformations of text images.

These challenges pose substantial obstacles to the implementation of OCR in our industrial context, with text erosion on iron surfaces being particularly problematic, even for manual inspection.

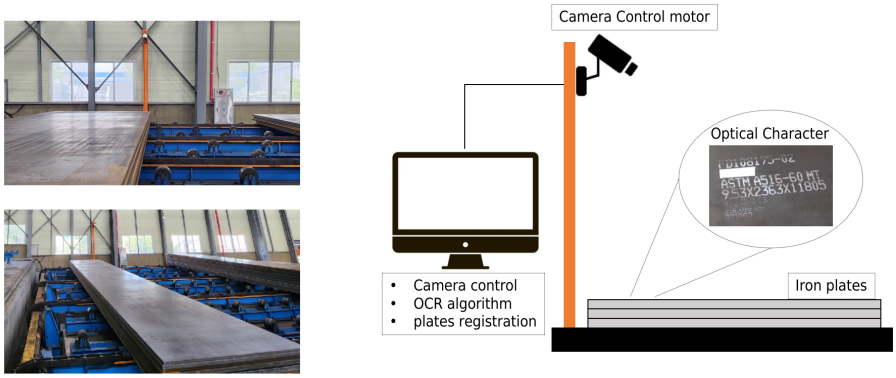


Figure 1. The designed autonomous iron plate registration system: (**a**) on-site photos, and (**b**) schematic of the iron plate registration system.

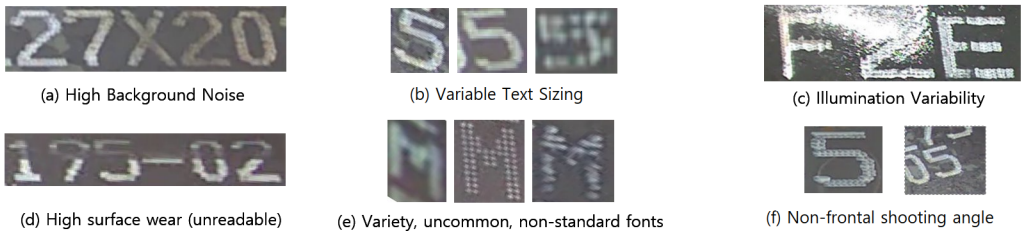


Figure 2. Example of low-quality images.

Due to the scarcity of adequate training datasets and the diversity of text fonts, primary applications in factory settings have traditionally relied on conventional computer vision-based OCR methodologies [12,16,18–20]. These include techniques such as template matching, edge detection, and Top Hat Filtering, primarily selected for their stability and reliability. However, these traditional methods necessitate manual calibration or configuration with auxiliary tools, which limits their applicability and poses a barrier to the advancement of fully autonomous factory systems. Furthermore, they exhibit limitations in extracting generalized features from diverse and variable conditions.

In contrast, recent advancements in Artificial Intelligence (AI), particularly with the emergence of deep learning (DL), have shown considerable potential in the domain of generalized feature extraction [21–23]. DL is adept at processing a wide range of font characters and enables automatic, nonlinear, and complex feature abstraction through numerous interconnected layers. This approach contrasts with traditional methods that rely on manually selected, predefined domain knowledge for optimal feature representation. However, existing research does not address the recognition of text on worn iron surfaces in industrial environments, a gap this paper aims to fill.

This study proposes an innovative approach that amalgamates camera control, deep learning, data augmentation, and synthetic image generation techniques. To our knowledge, this is the inaugural research effort specifically targeting text recognition on iron plates within an industrial context.

Our system architecture encompasses four stages, each contributing to the accuracy of the text recognition process. The on-site environment is illustrated in Figure 1a, and its schematic is depicted in Figure 1b. The architecture of our designed industrial OCR system is shown in Figure 3. Firstly, a text region detection network detects the text area, then controls the camera to rotate and zoom in to the text image region. Secondly, the detected text region undergoes segmentation line by line via a text segmentation network, followed by warping each line into straight rectangular patches for enhanced clarity. Thirdly, the text recognition network, composed of a vision text recognition model and a language network, recognizes each character in lines. Finally, the recognition results are registered into the iron plate ID registration system.

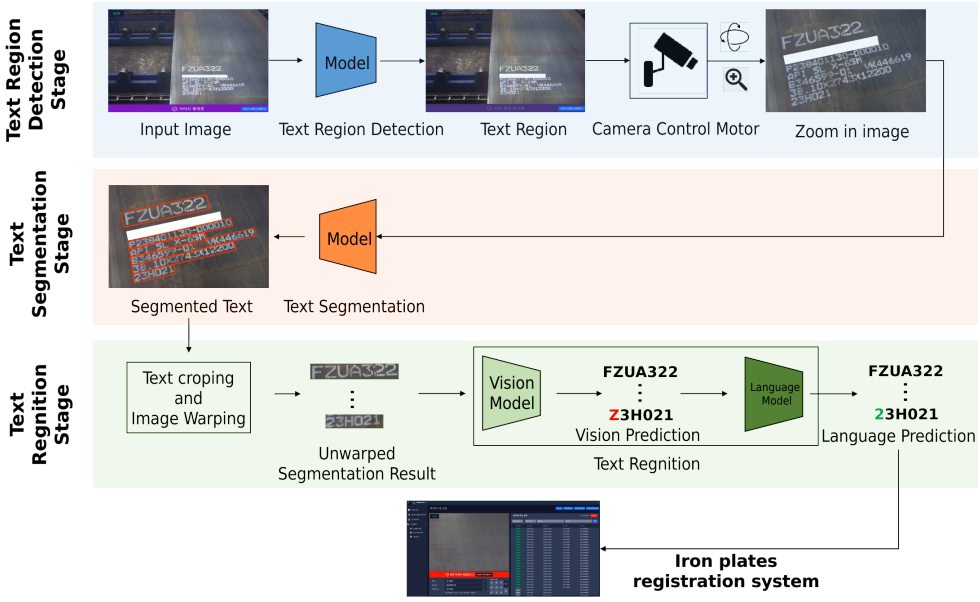


Figure 3. The architecture of our designed industrial OCR system. Text highlighted in red within the vision prediction denotes an incorrect prediction, whereas text highlighted in green within the language prediction indicates a correct prediction.

The overarching goal of this research is to develop a robust OCR system, tailored for industrial settings with limited data availability, and capable of delivering commercial-grade performance. This system encompasses the entire OCR pipeline, including data acquisition, the OCR neural network model, and a Graphical User Interface (GUI) for seamless integration into existing industrial processes.

To mitigate the challenge of insufficient training data, synthetic image generation was employed to enhance recognition accuracy. Moreover, we have implemented robust online data augmentation strategies, designed to replicate real-world conditions. These strategies include 'Random Rain' to simulate surface wear, 'Lightening Specific Areas' to mimic lighting effects, and 'Darkening Specific Areas' to simulate obstructions and shadows. The methodologies of the synthetic image generation and data augmentation will be expounded upon in Section 4.

In conclusion, this study makes significant contributions to the field of on-site industrial OCR through three main achievements:

1. This study develops a comprehensive on-site industrial OCR system tailored for the autonomous registration of iron plates. This system encompasses processing from the underlying algorithm to the user interface, offering a holistic solution that addresses the unique challenges of the industrial environment.
2. This study introduces a novel pre-processing methodology that includes the synthetic image generation (SIG) and strong data augmentation (SDA) techniques. These two technologies effectively mitigate the challenges posed by data scarcity in on-site industrial settings, enhancing the robustness and accuracy of the OCR system.
3. The designed autonomous OCR system significantly improves the efficiency of iron plate registration processes, leading to substantial reductions in both time and labor costs for factories. These outcomes not only showcase the practical benefits of our system but also highlight its potential to transform on-site industrial operations.

The structure of this article is as follows: Section 2 revisits the literature on OCR technology and its industrial applications. Section 3 delves into the proposed AI-based OCR system for industrial use, detailing OCR image acquisition, text segmentation, and text recognition models. Section 4 details the proposed pre-processing strategy, including synthetic image generation and data augmentation technologies. Section 5 presents the

experimental results, dataset and implementation details, comparative analyses with other models, and overall system performance. The paper concludes with Section 6.

2. Literature Review and Hypothesis Development

The Industrial OCR system, as delineated in this study, encapsulates the process of extracting text from images and transmuting optical characters into a digital format, subsequently archiving them in digital databases. This section aims to furnish a comprehensive overview of the overarching research landscape. Furthermore, based on the insights garnered from the literature review, this study posits two hypotheses.

2.1. Literature Review

2.1.1. General OCR

The domain of OCR in general scenarios, as delineated in the existing literature [24], encompasses a spectrum of tasks. These range from recognizing handwritten or printed documents to processing complex documents like invoices and bank statements, and even interpreting CAPTCHA systems [25,26]. Historically, early OCR techniques primarily focused on identifying individual characters or components, subsequently assembling them into words through the application of traditional computer vision approaches [18]. Contemporary research works, particularly those harnessing deep learning methodologies, tend to bifurcate the OCR process into distinct phases of text detection, text segmentation and text recognition.

In OCR, text detection methods can be bifurcated based on the granularity of target prediction into regression-based [27–29] and segmentation-based methods [21,23,30–32]. Regression-based techniques seek to demarcate text areas through approximate bounding boxes. Although akin to general object detection methods, they are often impeded by limitations in accurately capturing irregular and multi-oriented texts. The use of straight rectangular or quadrangular bounding boxes, prevalent in regression-based methods, frequently results in inaccuracies when dealing with texts of arbitrary shapes.

In contrast, segmentation-based methods exhibit enhanced aptitude in handling texts that are irregular, arbitrarily oriented, or multi-oriented. However, these methods typically require sophisticated post-processing algorithms to achieve optimal efficacy. For instance, Zhang et al. [30] utilized semantic segmentation combined with MSER-based algorithms for multi-oriented text detection, while Xue et al. [33] employed text borders to differentiate individual text instances.

A notable challenge in segmentation-based approaches is the risk of false detections, particularly when two text instances are in close proximity. Addressing this, PSENet [23] introduced the concept of progressive scale expansion, effectively segmenting text instances with varying scale kernels to separate closely situated lines. Furthering this innovation, DBNet++ [21] developed the Differentiable Binarization (DB) module, which integrates the binarization process into the segmentation network. This integration allows for the binarization process to be differentiable and optimized in tandem with the segmentation network, thus obviating the need for separate post-processing. Our industrial OCR system adopts DBNet++ to enhance text recognition efficacy.

Language-free text recognition methods [34–36], which primarily leverage visual features without accounting for character relationships, tend to underperform in low-quality image scenarios. Recognizing that text contains rich linguistic information, recent studies [22,37–39] have shifted focus towards language modeling. This approach has led to marginal improvements in OCR accuracy. Given the nature of our task, which is compounded by noisy inputs such as blurred and occluded text (as illustrated in Figure 2), we have selected to adopt a language-based methodology as the foundational model for our text recognition endeavors. The reason for the selection of language model is explained in Section 3.4.

2.1.2. Industrial OCR

Academic research tends to research OCR as two different stages, character segmentation and character recognition, as mentioned above. Academic research tends to improve character segmentation performance or character recognition performance separately.

However, in real-world applications, only using one network cannot cover the industrial application. The process of industrial OCR is composited by at least four steps: image acquisition, character segmentation, character recognition, and post-processing [11–17,19]. At the mean, the performance of these steps is inseparable because the output of one step is the input of the next step. Specifically, image acquisition, inclusive of pre-processing techniques, is designed to optimize image quality, thereby facilitating more effective character segmentation and recognition.

This paper introduces a comprehensive, on-site industrial OCR system tailored to these requirements. It encompasses image acquisition, character segmentation, character recognition, and iron plate ID registration. Our approach to image acquisition includes a text region detection network, which specifically focuses on obtaining a clearer capture of the text region, thereby enhancing the efficacy of the subsequent steps.

2.2. Hypothesis Development

Drawing on insights from our comprehensive literature review, this study proposes two hypotheses aimed at addressing key challenges in manufacturing settings through technological advancements.

Research Question 1: What are the potential benefits of developing and implementing a specialized and autonomous Optical Character Recognition (OCR) system for iron plates within industrial manufacturing environments?

Hypothesis 1: We hypothesize that by implementing a comprehensive OCR system designed for iron plate registration, it will lead to marked enhancements in operational efficiency and accuracy, thereby reducing manual labor and error rates [1–8]. Devasena et al. [5] highlight the growing importance of AI technologies in enhancing quality inspection processes across manufacturing sectors. This reference supports our assertion that AI-driven solutions, including advanced OCR systems, are pivotal in addressing the automation and efficiency needs of modern manufacturing environments. Furthermore, Kovvuri et al. [1–4,6] discuss the integration of AI in industrial production for improved efficiency and accuracy, further validating our research direction and the potential impact of our OCR system on manufacturing operations.

Research Question 2: How can the implementation of our designed synthetic image generation and strong data augmentation techniques mitigate the challenges posed by data scarcity in training OCR models for industrial applications?

Hypothesis 2: We expect that our innovative approach to synthetic image generation and data augmentation will effectively address the issue of data scarcity, enabling the OCR system to achieve high levels of accuracy in text recognition on iron plates, even under variable industrial conditions. The experimental results in Tables 1 and 2 highlight the enhancement achieved through our novel data pre-processing strategy, including synthetic image generation and strong data augmentation.

Table 1. Comparison among text segmentation model with our data-preprocessing strategy. **SIG**: synthetic image generation, **SDA**: strong data augmentation.

Model	Our Data Pre-Processing Strategy		Recall	Precision	F1 Score
	SIG	SDA			
DBNet [21]	✗	✗	0.3325	0.4961	0.3981
	✓	✗	0.3782	0.6257	0.4714
	✓	✓	0.3818	0.6592	**0.4836**

Table 1. *Cont.*

Model	Our Data Pre-Processing Strategy		Recall	Precision	F1 Score
	SIG	SDA			
Mask-RCNN [40]	✗	✗	0.4571	0.7558	0.5668
	✓	✗	0.4859	0.7256	0.5820
	✓	✓	0.5143	0.6828	**0.5867**
FCENet [41]	✗	✗	0.5013	0.7539	0.6022
	✓	✗	0.5841	0.7234	0.6463
	✓	✓	0.6130	0.6921	**0.6501**
DBNet++ [21]	✗	✗	0.5792	0.8577	0.6915
	✓	✗	0.7214	0.7532	0.7370
	✓	✓	0.7481	0.8067	**0.7763**

Table 2. Comparison among text recognition model with our data-preprocessing strategy. **SIG**: synthetic image generation, **SDA**: strong data augmentation.

Model	Our Data Pre-Processing Strategy		Recall	Precision	F1 Score	Accuracy
	SIG	SDA				
ASTER [42]	✗	✗	0.740	0.831	0.783	0.487
	✓	✗	0.826	0.876	0.850	0.601
	✓	✓	0.869	0.902	0.885	**0.636**
NTRT [43]	✗	✗	0.798	0.801	0.799	0.460
	✓	✗	0.826	0.897	0.860	0.611
	✓	✓	0.834	0.910	0.870	**0.623**
SATRN [44]	✗	✗	0.795	0.711	0.751	0.443
	✓	✗	0.843	0.879	0.860	0.625
	✓	✓	0.870	0.899	0.884	**0.659**
ABINet [22]	✗	✗	0.990	0.990	0.990	0.921
	✓	✗	0.990	0.991	0.990	0.926
	✓	✓	0.993	0.991	0.992	**0.928**

3. Proposed Industrial AI-Based OCR System

The proposed setup, designed for replication on client sites, consists of hardware components such as a computer (without a GPU), monitor, and cameras. Utilizing a CPU, as opposed to a GPU, offers a more cost-effective solution. A camera is strategically placed above the iron plates. Figure 1a and Figure 1b respectively depict the on-site photographs and the schematic of the iron plates registration system. The architecture of the system, depicted in Figure 3, encompasses four principal steps:

Step 1: OCR Image Acquisition by Text Region Detection Stage. This initial step identifies text regions using a specialized text region detection network. Following that, the camera is precisely adjusted to rotate along the x and y axes and zoom in to ensure the clarity of the images for the subsequent OCR task.

Step 2: Text Segmentation Stage. Once the text region is detected, text region are segmented line by line using a text segmentation network.

Step 3: Text Recognition Stage. Each line of text is transformed into rectangular patches through warping. Subsequently, a text recognition network is utilized to identify each character within these lines.

Step 4: Iron Plates Information Registration. The registration involves integrating the recognized text into the iron plate ID registration system. This step is crucial for enabling the autonomous registration and tracking of iron plates, streamlining the process of managing and identifying iron plates.

3.1. OCR Image Acquisition

Due to the varying lengths of iron plates and the arbitrary positioning of text on them, the size of the text region is not constant, leading to variable text resolution. Additionally, when the text region is distant, it tends to be smaller, potentially resulting in more background noise in the capture. This variability in text region size and the potential for increased background noise are key reasons why we opt not to use raw images (directly captured by the camera) for OCR. Instead, we employ a text region detection network to identify the Region of Interest (ROI), referred to as the text region, to perform OCR image acquisition.

The text region detection network locates the ROI, enabling the camera to rotate in the x–y direction and zoom in on the detected text area. This approach ensures that the characters are captured as clearly and closely as possible. The detected text region denoted as R_{TRD} is defined by its boundaries: the left boundary O_{x1}, right boundary O_{x2}, top boundary O_{y1}, and bottom boundary O_{y2}. The pipeline for the OCR image acquisition step is illustrated in Figure 3.

The camera's field of view is set after capturing an initial image. The computation of the camera rotation angles $\theta x, \theta y$ is based on the center points of the image and the object (text region). These center points are derived from the image dimensions I_{cx}, I_{cx}, and the detected text region R_{TRD}. The center of the image I_{cx}, I_{cx} and the center of the object O_{cx}, O_{cx} are calculated as follows:

$$I_{cx}, I_{cx} = \frac{I_{width}}{2}, \frac{I_{height}}{2} \tag{1}$$

$$O_{cx}, O_{cx} = O_{x1} + \frac{O_{x2} - O_{x1}}{2}, O_{y1} + \frac{O_{y2} - O_{y1}}{2} \tag{2}$$

Then , the horizontal and vertical offsets $\Delta X, \Delta Y$, representing the distance of the object's center from the image's center, are calculated as:

$$\Delta X, \Delta Y = O_{cx} - I_{cx}, O_{cy} - I_{cy} \tag{3}$$

The camera's rotation angles are then computed using these offsets and the camera's focal length f:

$$\theta_x, \theta_y = arctan(\frac{\Delta Y}{f}), arctan(\frac{\Delta X}{f}) \tag{4}$$

The zoom factor Z is determined based on the size of the text region and the image size, adjusted by a magnification factor k to ensure the entire text region is captured after zooming:

$$z = min(\frac{(O_{x2} - O_{x1}) \times k}{I_{width}}, \frac{(O_{y2} - O_{y1}) \times k}{I_{hight}}) \tag{5}$$

This approach allows for the precise positioning and zooming of the camera to capture the text region effectively.

3.2. Text Segmentation

For text segmentation, the system employs a Differentiable Binarization Network (DBNet) as detailed in the cited work [21]. The DBNet architecture, as shown in Figure 4, is designed to extract multi-scale features from an image of a text region using an Adaptive Feature-Pyramid Backbone Network [45,46]. The process involves predicting a probability map and a threshold map. Then, a binary map is generated by binarizing each pixel in the

probability map using the corresponding values from the threshold map. The final step in the DBNet process is the formation of bounding boxes (segmentation result) from the binary map, which is achieved through a specific box formation process.

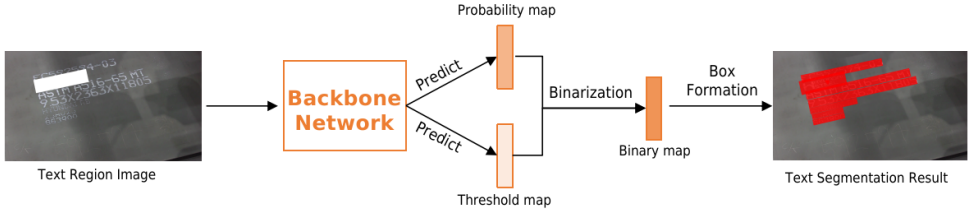


Figure 4. The architecture of the text segmentation network.

3.3. Segmented Text Image Warping

The presence of various orientations in zoomed-in images, often resulting from camera rotation and perspective distortions as illustrated in Figure 3, poses significant challenges for text recognition algorithms. Typically, these algorithms are more effective with horizontally aligned text because most of them are trained on public datasets, where the majority of the text is aligned horizontally.

To enhance the performance of text recognition in our system, it is essential to transform the segmented text results into horizontally aligned rectangular patches. This transformation is critical for improving text recognition accuracy, as it adjusts the text to a format more suitable for the pre-trained model.

Outlined in Algorithm 1 are the steps for transforming segmented text results into horizontally aligned (straight) rectangular patches. This is achieved using a warping perspective algorithm, which effectively adjusts the text's orientation and perspective. Once the text is horizontally aligned, it is extracted as a rectangular patch. This patch is then ready for input into the text recognition algorithm.

The text segmentation result is represented as multiline quadrilaterals with various angles as illustrated in Figure 3. The coordinates of the segmented text region are denoted by *src*, which include the points $[(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)]$. These points represent the most upper left, upper right, bottom right, and bottom left of the segmentation result, respectively.

The target shape for the warped text segmentation result is a horizontally aligned rectangle and is set by us, with dimensions $W_{\text{warp}} \times H_{\text{warp}}$. The coordinates for this target shape, denoted as *dest*, are $(0, 0)$, $(W_{\text{warp}}, 0)$, $(W_{\text{warp}}, H_{\text{warp}})$, and $(0, H_{\text{warp}})$, corresponding to the most upper left, upper right, bottom right, and bottom left of the warped segmentation result.

A 3×3 perspective transformation matrix M is calculated using *src* and *dest* in step 2. M is then used to warp the text segmentation result from the original image into horizontally aligned rectangular patches in step 3. Example images of the text segmentation results and their warped counterparts are shown in the experiment section. .

Algorithm 1 Text segmentation image warping.

Require: Z: Zoom in image
Require: Text segmentation result *src*: [(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)]
Require: W_{warp}, H_{warp}: Width and height of the warped image
Ensure: U
Step 1: $dest \leftarrow [(0,0), (W_{warp}, 0), (W_{warp}, H_{warp}), (0, H_{warp})]$
Step 2: transformation matrix $M \leftarrow$ cv2.getPerspectiveTransform$(src, dest)$
Step 3: warped text segmentation result $U \leftarrow$ cv2.warpPerspective$(Z, M, (W_{warp}, H_{warp}))$

As previously mentioned, iron plates often present low-quality data challenges, such as wear, light and sunshine reflection, and occlusion, which hinder the effectiveness of vision-based text recognition methods in these challenging cases. Recent research [22] has

demonstrated that incorporating linguistic information can significantly enhance recognition accuracy in low-quality images.

3.4. Language-Based Text Recognition Network

Building on this concept, our research implements a language-based character recognition approach known as the Autonomous, Bidirectional, and Iterative Network (ABINet) [22]. Figure 3 shows the architecture of the text recognition network.

ABINet combines a vision-based text recognition model with a language model. Initially, the vision model predicts a preliminary recognition result, which the language model subsequently refines to produce the final outcome. ABINet comprises three main components:

1. Autonomous Strategy: This strategy enables the language model to operate independently of the vision model, allowing for its replacement. In other words, the language model can be trained separately and then serve as a spelling corrector, enhancing the output of the vision model.
2. Bidirectional Representation: Contrary to models that process character sequences in a unidirectional (left-to-right) manner, ABINet adopts a bidirectional approach, analyzing language features from both left to right, and right to left. This bidirectional representation is crucial for our OCR tasks in industrial products, where the position of characters often conveys specific meanings with patterns. For instance, certain positions might indicate the production date using only numbers, while others could denote the quality level of the material with English letters. The effectiveness of ABINet in these scenarios where character positioning is meaningful highlights its superiority for industrial product OCR tasks. The experimental results presented in Table 2 also support this assertion.
3. Iterative Correction: ABINet introduces an iterative correction mechanism to address the challenges posed by noisy inputs, such as blurred or occluded text. The language model iteratively refines the vision model's predictions, making it particularly suitable for our application where text quality is compromised by factors like poor lighting or wear and tear as discussed in Section 1 and illustrated in Figure 2.

In conclusion, the selection of ABINet for our text recognition network was based on its bidirectional representation and iterative correction capabilities, which are particularly advantageous for our industrial OCR applications.

4. Proposed Data Pre-Processing Strategy for Industrial OCR System

DL-based methods, such as DBNet++ [21] and ABINet [22], typically require extensive data for training. However, in industrial contexts, acquiring substantial amounts of data can be challenging. The scarcity of training data presents a significant obstacle for AI-based applications in industrial settings. To address this issue of data deficiency, we have implemented two key strategies: the generation of synthetic images and the adoption of a strong data augmentation strategy. These approaches aim to enhance the volume and diversity of data available for training the system, thereby mitigating the limitations posed by the lack of real-world industrial data.

4.1. Synthetic Image Generation

Our training dataset includes both actual images and synthetically generated non-real images. The aim is to create synthetic images that mirror real-world scenarios, thereby expanding the volume and variety of data available for effectively training our model. The process of generating these synthetic images is depicted in Figure 5.

The first step in this process involves collecting background images from actual real-world photographs. To increase the diversity of the synthetic images, a random pre-processing step is applied. This step includes operations such as random cropping and random flipping, which add variability to the images.

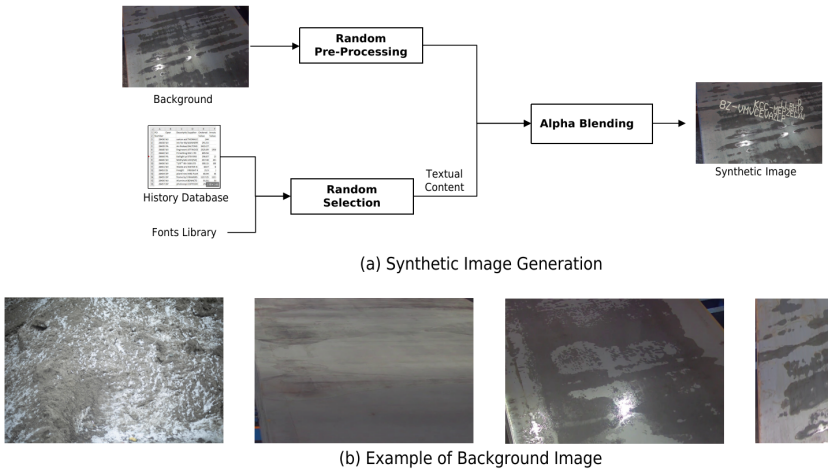


Figure 5. Synthetic image generation.

The textual content for these synthetic images is randomly selected from a 'history database'. This database consists of data that have been manually recorded by workers in the past. It serves as a rich source of varied text that can be superimposed on the background images.

To merge the textual content with the background images, the Alpha Blending method is utilized. During this process, several parameters are randomly adjusted to make the synthetic images more realistic. These adjustments include random positioning of the text, varying text angles between $(-30, 30)$, and applying perspective transformations. These manipulations are specifically designed to mimic non-frontal shooting angles, which are common in real-world scenarios.

As depicted in Figure 5b, examples of background images are presented, showcasing the diversity in the types of images used for generating synthetic data. A key advantage of these synthetic images is that they come with accurate label information, which is embedded during the generation process. This feature eliminates the need for human annotation, significantly streamlining the data preparation phase for model training. In this particular instance, a total of 1000 synthetic images were generated. The inclusion of these images in the training dataset enriches the variety and volume of data available.

4.2. Strong Data Augmentation

To enhance the model's performance in real-world conditions, we have developed a specific data augmentation strategy that closely replicates on-site scenarios. As highlighted in the Introduction Section 1 and illustrated in Figure 2, the data for training are low quality. To address this challenge and generate more complex training images, various data augmentation techniques are employed. Our strategy includes both simple and common data augmentation techniques as well as more tailored and robust methods. The simpler techniques encompass the following:

- Blurring: to mimic variable text resolution that might arise from varying distances between the camera and the target text;
- Random Brightness: implemented to simulate changes in the illumination conditions.

However, we also confront the challenge of high background noise, which is a prevalent issue in real-world scenarios. To tackle this, we have incorporated tailored and stronger data augmentation techniques as given below:

- Random Rain: this technique is used to simulate the appearance of water stains on the images.
- Lightening Specific Regions: this technique mimics the effects of light or sun reflections that can interfere with text visibility.

- Darkening Specific Areas: this is applied to represent shadows or occlusions that can obscure parts of the text.

Figure 6 shows the application of the above techniques. The upper row of images in the figure displays original real-world images that are affected by water stains, light reflections, and shadows. The lower row, in contrast, shows images after the application of our strong data augmentation techniques: Random Rain, Lightening, and Darkening. These augmented images provide a more challenging and varied dataset, preparing the model to handle a wide range of visual disturbances that it may encounter in iron plate practical applications. This approach ensures the robustness of the model in diverse and challenging industrial iron plate recognition applications.

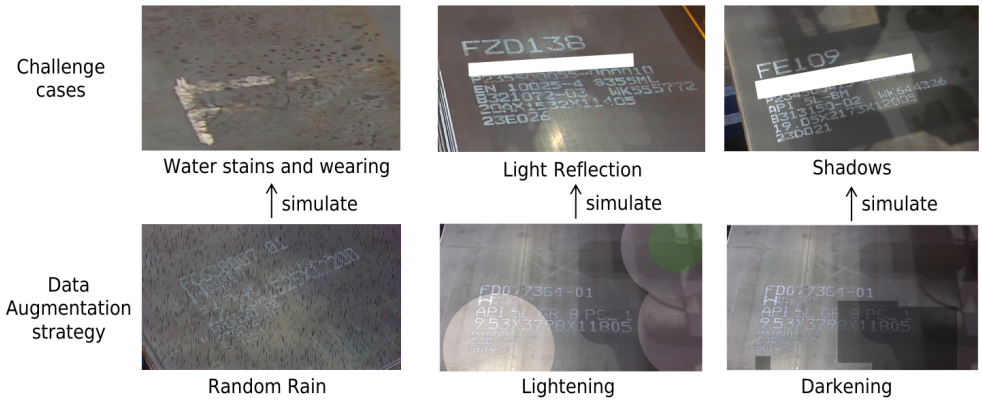


Figure 6. Example images of data augmentation.

It is important to note that both synthetic image generation and strong data augmentation techniques are exclusively utilized during the training phase of the model.

5. Experiment

5.1. Datasets and Implementation Details

Our training datasets consist of both publicly available datasets and real-world datasets. The public datasets are utilized exclusively for pretraining three models: the text region detection model, the text segmentation model, and the text recognition model. The real-world dataset comprises images captured by on-site cameras, totaling 212 images. These are distributed as follows: 145 images for training, 45 images for validation, and 22 images for testing. Additionally, 1000 synthetic images are generated and used solely for training purposes. Therefore, the total count for training is 1145 images, with 45 images for validation and 22 for testing. Both real and synthetic images are employed in the training of the three models.

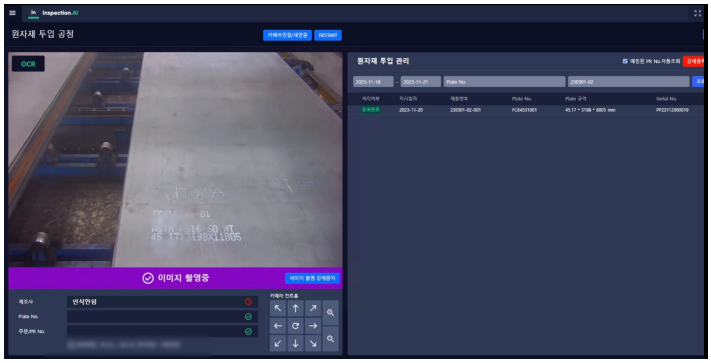
The training of these models is conducted on a system equipped with an Intel i5-12400F CPU and an NVIDIA 3080Ti 12G GPU, running CUDA version 12.0. For cost efficiency, the models are trained using GPU but run on CPU-only computers, which is more economical than using a GPU. The overall system speed, from image acquisition to iron plate ID matching, is approximately 1.4 s per image. The raw images captured by the camera are of size 1280 × 720. All other hyperparameters and the optimizer, not mentioned below, are set to their default values as specified in MMOCR [47].

For text region detection, YOLOv5 is chosen due to its high performance in both accuracy and speed. The model is fine-tuned for 500 epochs on the real-world dataset, using a network pre-trained on the COCO [48] general object detection dataset. Images are resized to 640 × 640 pixels, with a training batch size of 4. The learning rate is set at 0.01.

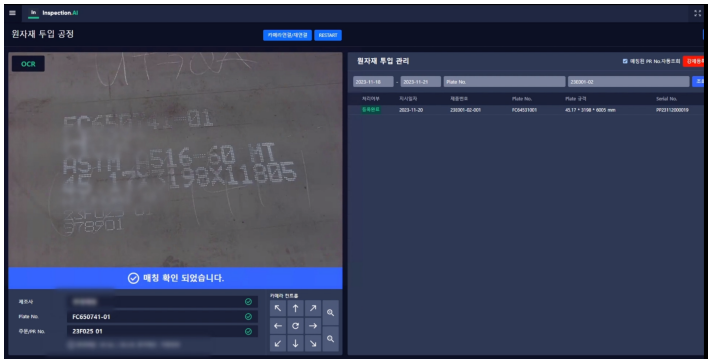
DBNet++ [21] is utilized for text segmentation. This model undergoes fine-tuning for 200 epochs on the real-world dataset, leveraging a network pre-trained on the ICDAR2015 [49]. The images are resized to 640 × 640 pixels, with a training batch size of 4. The learning rate is maintained at 0.01.

For text recognition, ABINet++ [22] is employed. The model is fine-tuned for 200 epochs on the real-world dataset, based on a network pre-trained on the ICDAR2015 [49]. The image size for this phase is 32×128 pixels, with the training batch size increased to 16. The learning rate is set at 0.01.

The GUI of our OCR system is illustrated in Figure 7. The software is designed to support the matching and registration of recognized product IDs. The system facilitates both autonomous and manual registration of iron plates. The left side of the GUI displays the real-time images of on-site iron plates captured by the camera. The bottom left window shows the recognized text. The right side of the interface is dedicated to matching and registering recognized product IDs, allowing users to check the status of these IDs. Additionally, the software includes several buttons to control the system and to register product IDs that were not successfully recognized. This interface is a crucial part of the OCR system, ensuring user-friendly interaction and efficient management of the OCR process.



(**a**)



(**b**)

Figure 7. Illustration of the GUI: (**a**) before registration, (**b**) after registration.

5.2. Comparison Result

This section validates the effectiveness of our proposed data pre-processing strategy, which includes synthetic image generation and strong data augmentation that can help both text segmentation models and text recognition models.

5.2.1. Text Segmentation Model

Table 1 showcases the results from a comparative analysis of different text segmentation models, highlighting the enhancement achieved through our novel data preprocessing strategy. By incorporating this strategy, we observed a uniform improvement in performance metrics across four models: DBNet [21], Mask-RCNN [40], FCENet [41], and DBNet++ [21]. Notably, the F1 scores increased from 0.3981 to 0.4836 for DBNet [21], from 0.5668 to 0.5867 for Mask-RCNN [40], from 0.6022 to 0.6501 for FCENet [41], and from 0.6915 to 0.7763 for DBNet++ [21]. This uniform enhancement underscores the robustness and efficacy of our approach across a range of models. Among them, DBNet++ outper-

formed the others, demonstrating superior efficacy, especially when augmented with our pre-processing strategy, despite a minor decrease in precision. Nevertheless, the gains in recall and F1 score suggest that the integration of the pre-processing strategy fortifies the model's robustness, effectively balancing the precision trade-off with significant overall improvements.

Furthermore, Table 1 demonstrates the stepwise enhancements brought about by the SIG and SDA techniques on real-world datasets. Initially, the deployment of SIG notably elevated the F1 scores to 0.4714 for DBNet [21], 0.5820 for Mask-RCNN [40], 0.6463 for FCENet [41], and 0.7370 for DBNet++ [21]. Further application of SDA contributed to an additional increase in these scores to 0.4836, 0.5867, 0.6501, and 0.7763, respectively. These advancements underscore the effectiveness of our methodologies in significantly improving model performance across various benchmarks.

Figure 8 illustrates the error analysis of the text segmentation network. This analysis encompasses both simple and complex scenarios. Simple cases, as shown in Figure 8a, feature a clear background. Conversely, complex cases are displayed in Figure 8b,c, characterized by challenges such as water stains, wear, light reflections, and shadows. The first row of images in Figure 8 represents the ground truth, while the second row depicts the segmentation outcomes from the baseline model. The third row presents the results of the baseline model with our proposed strong data augmentation strategy.

The performance of the baseline model, depicted in the second row, deteriorates significantly in the complex scenarios of Figure 8b,c. These results highlight the detrimental impact of challenging factors like water stains, wear, light reflections, and shadows on performance. In other words, the result images in the second row illustrate the errors introduced by the challenging factors. To address these challenges, we introduced a strong data augmentation strategy that includes a random rain method to simulate water stains and wear, a lighting method to replicate light reflections, and darkening effects to mimic shadows. The outcomes of the designed strong data augmentation strategy, shown in the third row, demonstrate notable improvements in model performance.

In conclusion, the empirical evidence presented in Table 1 unequivocally establishes the independent and cumulative impact of our proposed synthetic image generation and strong data augmentation techniques. These methodologies not only enhance model performance significantly but also demonstrate a robust adaptability to real-world datasets. The consistent improvement across different models underscores the effectiveness of our approaches, validating their utility as standalone enhancements as well as in combination. Such findings bolster the argument for integrating these strategies into the preprocessing pipeline, offering a substantial leap forward in the field of text segmentation.

5.2.2. Text Recognition Model

Similar to Table 1, Table 2 further presents a comparative analysis of the text recognition models' performance, with and without our data-preprocessing strategy, which includes SIG and SDA. As in Section 5.2.2, we evaluated four models—ASTER [42], NTRT [43], SATRN [44], and ABINet [22]—to demonstrate our preprocessing strategy's effectiveness across different systems. The analysis revealed two key findings:

Firstly, ABINet outperforms the other models, achieving higher accuracy rates than ASTER by 0.434, NTRT by 0.461, and SATRN by 0.478. This superior performance is attributed to the ABINet advanced language model, which excels in recognizing the structured and meaningful placement of characters in industrial settings. As we mentioned in Section 3.4, specific positions of characters in iron plates denote the production date or material quality of iron plates, using a constrained set of numbers or English letters. The ABINet design effectively captures and interprets these patterns.

Secondly, implementing our proposed data pre-processing strategy significantly enhances the text recognition accuracy across all models, including ABINet. The introduction of SIG improved the ABINet word accuracy from 0.921 to 0.926, while SDA further boosted it to 0.928. Similarly, SIG enhanced the ASTER word accuracy from 0.487 to 0.601, with SDA

elevating it to 0.636. Notable improvements were also observed in NTRT and SATRN; following our preprocessing strategy, their performance results were elevated to 0.623 and 0.659, respectively.

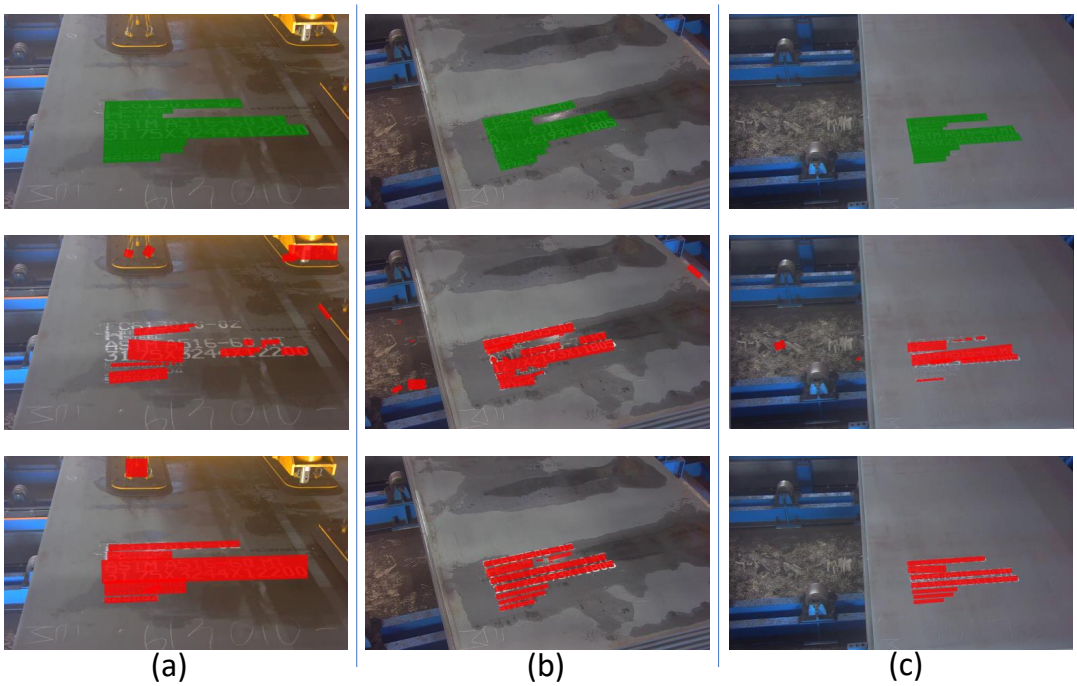


Figure 8. Error analysis of text segmentation network. Three cases (**a**–**c**) are shown.

In conclusion, the results in Table 2 underscore the exceptional capability of ABINet in industrial OCR tasks, where the positional meaning of characters is critical. Moreover, our data-preprocessing strategy of SIG and SDA consistently elevates the performance of text recognition models on real-world datasets, affirming its effectiveness.

5.3. System Result

In this subsection, we discuss the qualitative outcomes from applying our developed OCR system to a real-world task involving industrial iron plates. The discussion is structured to highlight each phase of the process, including the detection of text regions, segmentation of text, and the recognition of text. Through this structured approach, we aim to illustrate the effectiveness and intricacies of each step within the OCR system.

Figure 9 displays the outcomes of text region detection and camera control in our OCR system. Two examples, each with a different input image, are presented in separate rows. The raw image, captured by the camera, is processed by the trained text region detector. The accurately detected text regions are denoted by green-colored boxes. Additionally, red-colored bounding boxes, which are enlarged versions of the green boxes using the magnification factor k as defined in Equation (5), are shown. These enlarged boxes ensure complete capture of the text region upon camera zoom-in. The zoomed-in images demonstrate the effectiveness of the text region detection model and the camera control motor.

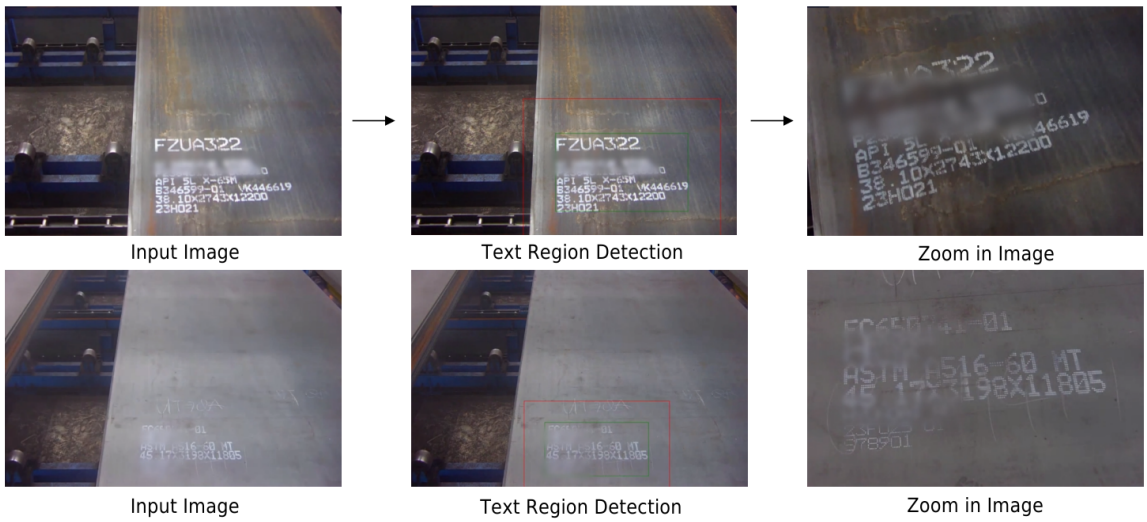
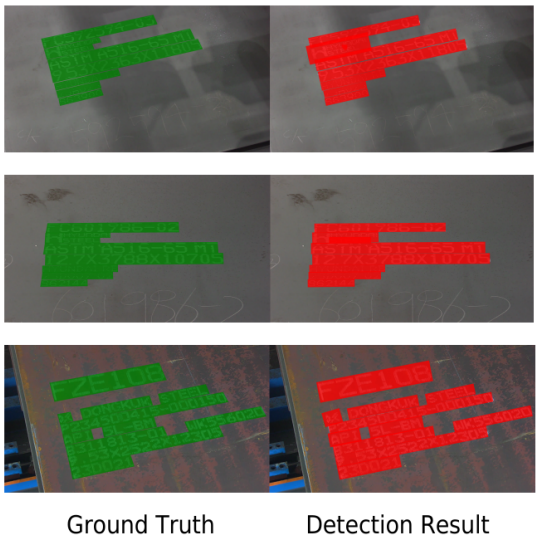


Figure 9. Result of text region detection.

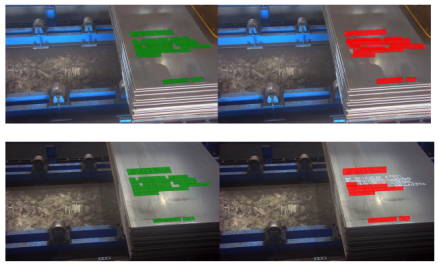
Figure 10a illustrates the results of text segmentation in our system. The figure demonstrates that our system is capable of accurately detecting text lines, even when they are tightly spaced or close to each other. Particularly, Figure 10b highlights the segmentation results obtained without prior application of text region detection. In this scenario, the text regions are smaller and do not yield clear images of text lines, leading to misalignment in the segmentation results. If some text regions cannot be segmented properly, it negatively impacts the subsequent text recognition process due to the low resolution of characters. The contrastive experimental results between Figure 10a,b also underline the importance of applying a text region detection network before segmenting text. Moreover, implementing text region detection as an initial step aids in removing unnecessary text, such as that located at the bottom of iron plates, thereby refining the segmentation process.

Figure 11 presents a comparative analysis of the character recognition results between (a) the baseline method and (b) the baseline method augmented with our data pre-processing strategy, including synthetic image generation and strong data augmentation techniques. The recognition results are displayed below each image. The ground truth is indicated on the left side. Incorrectly recognized characters by the baseline method are marked in red, whereas characters correctly recognized by our method (but not by the baseline) are highlighted in green. While both the baseline and our method accurately predict the first image, the baseline method incorrectly identifies "A" as "H" in the second image and "V" as "M". This confusion arises due to the similar shapes of these characters and can occur when the dataset is insufficient. However, our proposed synthetic image generation approach increases the dataset's volume, covering such confusion cases and leading to the correct recognition of these challenging characters.

These results further demonstrate the effectiveness of our method in handling complex real-world scenarios, specifically in the challenging context of industrial iron plate inspection. By providing a step-by-step visual breakdown, Section 5.3 offers a comprehensive view of our method's capabilities in accurately detecting and recognizing text on iron plates, highlighting its robustness and practical applicability in industrial applications.



(**a**)



(**b**)

Figure 10. Comparison result of text segmentation (**a**) with our designed text region detection network and (**b**) without the text region detection network.

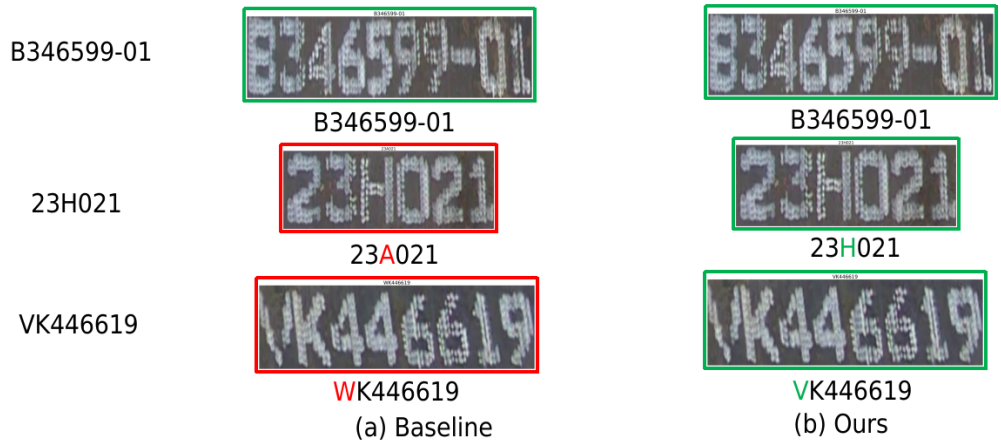


Figure 11. Comparison result of character recognition between the (**a**) baseline method and (**b**) the baseline method with our designed synthetic image generation and strong data augmentation.

6. Conclusions

This research paper has delved into the development of a specialized on-site industrial Optical Character Recognition (OCR) system, primarily focused on deciphering textual content on iron plates. In the context of sustainable development, this system presents several contributions to the manufacturing industry. Firstly, the introduction of an automated, efficient OCR process for iron plate registration aligns with the principles of sustainable manufacturing. By significantly reducing manual labor and processing times, the system

contributes to a reduction in energy usage and operational costs. This efficiency not only bolsters the economic sustainability of manufacturing operations but also minimizes the environmental footprint by reducing the energy intensity per unit of production. Moreover, the system's ability to accurately and rapidly process information contributes to the optimization of supply chain and inventory management. This leads to a more judicious use of resources, decreasing waste and enhancing material efficiency—key tenets of sustainable manufacturing. The precise tracking and data management enabled by our OCR system can aid in better resource allocation and waste reduction, both of which are crucial for sustainable industrial practices.

Another significant aspect is the potential reduction in error rates compared to manual processes. This accuracy not only improves operational efficiency but also reduces the likelihood of resource misuse and waste generation, furthering the cause of environmental sustainability. Furthermore, the technology developed in this research can be adapted and extended to other industrial applications, potentially leading to broader impacts on sustainability across various manufacturing sectors. By automating and optimizing processes that were traditionally labor-intensive and error-prone, this technology paves the way for more sustainable manufacturing practices industry-wide.

Finally, this research contributes to the knowledge base of sustainable industrial practices, offering insights and a practical solution that can inspire future innovations aimed at enhancing sustainability in the manufacturing sector. It embodies an intersection of technological advancement and sustainable development, demonstrating how digital technologies can be leveraged to achieve greater efficiency and environmental responsibility in industrial settings. In conclusion, the development of this OCR system is not just a technological advancement but also a stride towards more sustainable manufacturing practices. It underscores the potential of integrating advanced technologies in industrial operations to foster economic, environmental, and operational sustainability.

Author Contributions: Methodology, Q.T.; Visualization, Y.L.; Supervision, H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ulsan City & Electronics and Telecommunications Research Institute (ETRI) grant funded by the Ulsan City [23AS1600, the development of intelligentization technology for the main industry for manufacturing innovation and human–mobile–space autonomous collaboration intelligence technology development in industrial sites].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The iron plate OCR dataset is available upon request. Please contact the corresponding author for data.

Conflicts of Interest: Author Q.T and Y.L was employed by the company INTERX. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Haseeb, M.; Hussain, H.I.; Ślusarczyk, B.; Jermsittiparsert, K. Industry 4.0: A Solution towards Technology Challenges of Sustainable Business Performance. *Soc. Sci.* **2019**, *8*, 154 . [CrossRef]
2. Sanchez, M.; Exposito, E.; Aguilar, J. Industry 4.0: Survey from a system integration perspective. *Int. J. Comput. Integr. Manuf.* **2020**, *33*, 1017–1041. [CrossRef]
3. Oztemel, E.; Gursev, S. Literature review of Industry 4.0 and related technologies. *J. Intell. Manuf.* **2018**, *31*, 127–182. [CrossRef]
4. Woschank, M.; Rauch, E.; Zsifkovits, H. A Review of Further Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics. *Sustainability* **2020**, *12*, 3760. [CrossRef]
5. Devasena, D.; Dharshan, Y.; Vivek, S.; Sharmila, B. AI-Based Quality Inspection of Industrial Products. In *Handbook of Research on Thrust Technologies Effect on Image Processing*; IGI Global: Hershey, PA, USA, 2023; pp. 116–134. [CrossRef]
6. Kovvuri, R.R.; Kaushik, A.; Yadav, S. Disruptive technologies for smart farming in developing countries: Tomato leaf disease recognition systems based on machine learning. *Electron. J. Inf. Syst. Dev. Ctries.* **2023**, *89*, e12276. [CrossRef]

7. Li, L.; Lv, M.; Jia, Z.; Ma, H. Sparse Representation-Based Multi-Focus Image Fusion Method via Local Energy in Shearlet Domain. *Sensors* **2023**, *23*, 2888. [CrossRef] [PubMed]
8. Zhang, X.; Li, W.; Gao, C.; Yang, Y.; Chang, K. Hyperspectral pathology image classification using dimension-driven multi-path attention residual network. *Expert Syst. Appl.* **2023**, *230*, 120615. [CrossRef]
9. Jung, H.; Rhee, J. Application of YOLO and ResNet in Heat Staking Process Inspection. *Sustainability* **2022**, *14*, 15892. [CrossRef]
10. Tang, Q.; Jung, H. Reliable Anomaly Detection and Localization System: Implications on Manufacturing Industry. *IEEE Access* **2023**, *11*, 114613–114622. [CrossRef]
11. Wang, X.; Li, Y.; Liu, J.; Zhang, J.; Du, X.; Liu, L.; Liu, Y. Intelligent Micron Optical Character Recognition of DFB Chip Using Deep Convolutional Neural Network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–9. [CrossRef]
12. Caldeira, T.; Ciarelli, P.M.; Neto, G.A. Industrial Optical Character Recognition System in Printing Quality Control of Hot-Rolled Coils Identification. *J. Control Autom. Electr. Syst.* **2020**, *31*, 108–118. [CrossRef]
13. Subedi, B.; Yunusov, J.; Gaybulayev, A.; Kim, T.H. Development of a Low-cost Industrial OCR System with an End-to-end Deep Learning Technology. *J. Embed. Syst. Appl.* **2020**, *15*, 51–60.
14. Cai, B. Deep learning Optical Character Recognition in PCB Dark Silk Recognition. *World J. Eng. Technol.* **2023**, *11*, 1–9. [CrossRef]
15. Zhang, C.; Liu, B.; Chen, Z.; Yan, J.; Liu, F.; Wang, Y.; Zhang, Q. A Machine Vision-Based Character Recognition System for Suspension Insulator Iron Caps. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–13. [CrossRef]
16. Kazmi, W.; Nabney, I.; Vogiatzis, G.; Rose, P.; Codd, A. An Efficient Industrial System for Vehicle Tyre (Tire) Detection and Text Recognition Using Deep Learning. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1264–1275. [CrossRef]
17. Paglinawan, C.C.; Caliolio, M.H.M.; Frias, J.B. Medicine Classification Using YOLOv4 and Tesseract OCR. In Proceedings of the 2023 15th International Conference on Computer and Automation Engineering (ICCAE), Sydney, Australia, 3–5 March, 2023; pp. 260–263.
18. Neumann, L.; Matas, J. Real-time scene text localization and recognition. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3538–3545.
19. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*, 3rd ed.; Pearson: London, UK, 2007
20. Yang, C.; Yang, Y. Improved local binary pattern for real scene optical character recognition. *Pattern Recognit. Lett.* **2017**, *100*, 14–21. [CrossRef]
21. Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; Bai, X. Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 919–931. [CrossRef] [PubMed]
22. Fang, S.; Mao, Z.; Xie, H.; Wang, Y.; Yan, C.; Zhang, Y. ABINet++: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Spotting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7123–7141. [CrossRef] [PubMed]
23. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, Long Beach, CA, USA, 15–20 June 2019; pp. 9336–9345.
24. Mudhsh, M.; Almodfer, R. Arabic Handwritten Alphanumeric Character Recognition Using Very Deep Neural Network. *IInformation* **2017**, *8*, 105. [CrossRef]
25. Mathew, A.; Kulkarni, A.; Antony, A.; Bharadwaj, S.; Bhalerao, S. DOCR-CAPTCHA: OCR Classifier based Deep Learning Technique for CAPTCHA Recognition. In Proceedings of the 2021 19th OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 16–18 December 2021; pp. 347–352.
26. Alsuhibany, S.A.; Parvez, M.T. Secure Arabic Handwritten CAPTCHA Generation Using OCR Operations. In Proceedings of the 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 23–26 October 2016; pp. 126–131.
27. Liao, M.; Shi, B.; Bai, X. Textboxes++: A single-shot oriented scene text detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
28. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. Textboxes: A fast text detector with a single deep neural network. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
29. He, P.; Huang, W.; He, T.; Zhu, Q.; Qiao, Y.; Li, X. Single shot text detector with regional attention. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, Venice, Italy, 22–29 October 2017; pp. 3047–3055.
30. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multioriented text detection with fully convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, Las Vegas, NV, USA, 27 June–30 June 2016.
31. Liao, M.; Lyu, P.; He, M.; Yao, C.; Wu, W.; Bai, X. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 532–548. [CrossRef] [PubMed]
32. Lyu, P.; Liao, M.; Yao, C.; Wu, W.; Bai, X. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In Proceedings of the European Conference on Computer Vision (ECCV), 2018, Munich, Germany, 8–14 September 2018; pp. 67–83.
33. Xue, C.; Lu, S.; Zhan, F. Accurate Scene Text Detection through Border Semantics Awareness and Bootstrapping. *arXiv* **2018**, arXiv:1807.03547 .

34. Graves, A.; Fernandez, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labeling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376
35. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2359–2367.
36. Wan, Z.; He, M.; Chen, H.; Bai, X.; Yao, C. Textscanner: Reading characters in order for robust scene text recognition. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), New York, NY, USA, 7–12 February 2020.
37. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep structured output learning for unconstrained text recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
38. Lee, C.-Y.; Osindero, S. Recursive recurrent nets with attention modeling for ocr in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27 June–30 June 2016; pp. 2231–2239.
39. Qiao, Z.; Zhou, Y.; Yang, D.; Zhou, Y.; Wang, W. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13528–13537.
40. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, Venice, Italy, 22–29 October 2017.
41. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3122–3130.
42. Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2035–2048. [CrossRef] [PubMed]
43. Sheng, F.; Chen, Z.; Xu, B. NRTR: A No-Recurrence Sequence-to-Sequence Model for Scene Text Recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 781–786.
44. Lee, J.; Park, S.; Baek, J.; Oh, S.J.; Kim, S.; Lee, H. On Recognizing Texts of Arbitrary Shapes with 2D Self-Attention. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 2326–2335.
45. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; =Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
46. Q. Tang, G. Cao and K. -H. Jo, Integrated Feature Pyramid Network With Feature Aggregation for Traffic Sign Detection. *IEEE Access* **2021**, *9*, 117784–117794. [CrossRef]
47. Kuang, Z.; Sun, H.; Li, Z.; Yue, X.; Lin, T.H.; Chen, J.; Wei, H.; Zhu, Y.; Gao, T.; Zhang, W.; et al. MMOCR: A Comprehensive Toolbox for Text Detection, Recognition and Understanding. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021.
48. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
49. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Gomez, L.; Robles, S.; Mas, J.; Fernandez, D.; Almazan, J.; Heras, L.P.d. ICDAR 2013 Robust Reading Competition. In Proceedings of the 12th International Conference of Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1115–1124.