Maximizing Recognition Reliability: TrOCR Outperforms PaddleOCR in Challenging Container Automation Environments

Jia Qing Cheok[1], Kim Chuan Lim[1*], Chong En Si[2]

[1]Centre for Telecommunication Research and Innovation (CeTRI), Fakulti Teknologi dan Kejuruteraan Elektronik dan Komputer, Universiti Teknikal Malaysia Melaka

[2]Faculty of Management, Multimedia University, Cyberjaya, Selangor, Malaysia

[*]Corresponding Author

DOI: https://dx.doi.org/10.47772/IJRISS.2025.91100629

Received: 11 December 2025; Accepted: 18 December 2025; Published: 29 December 2025

ABSTRACT

Container automation systems have become increasingly important in response to the rapid growth of global trade and the need for efficient logistics. Previous research lacked a systematic comparison of advanced OCR models (PaddleOCR and TrOCR) integrated with reliable text detection (YOLO) to determine the optimal balance of speed and high accuracy under real-world port conditions. This study developed an automated pipeline combining the YOLOv10 object detector for text region localization with fine-tuned PaddleOCR and TrOCR models for recognition. Evaluation was conducted on a test set of 173 real-world images from an actual port terminal gate deployment after training on 8,899 augmented images. YOLOv10 achieved strong detection performance, recording a mean Average Precision (mAP) of 94.7% and an average Intersection over Union (IoU) of 0.87. TrOCR consistently demonstrated superior recognition accuracy, achieving 98.73% exact match for ISO codes and 71.17% for container numbers, exceeding PaddleOCR (97.42% and 70.14%). However, PaddleOCR was significantly faster (up to 18.35 FPS for ISO codes) compared to TrOCR (7.93 FPS). The integrated YOLOv10 with TrOCR pipeline is recommended for reliable, high-precision text recognition, advancing automated port logistics through a scalable, AI-powered solution that prioritizes accuracy in challenging real-world scenarios.

Keywords: Deep Learning, Container Text Detection System, Container Text Recognition.

INTRODUCTION

Container automation systems are vital infrastructure in global logistics, demanding highly efficient operations where the real-time identification of container identifiers (such as container numbers and ISO codes) is paramount. Developed nations like the Netherlands [1], China [2],[5] and Australia [3],[4] leverage sophisticated Optical Character Recognition (OCR) systems integrated with AI to manage container flow and achieve high throughput. The operational environment of ports, particularly in regions like Malaysia, presents unique challenges to these systems, including inconsistent container markings, complex backgrounds, poor lighting, text rotation, and text degradation (*e.g.*, rust or fading), which severely hinder reliable automated text extraction.

Effective container recognition with conventional surveillance camera relies on a two-stage deep learning pipeline: robust text detection followed by high-fidelity text recognition (see Figure 1). For the detection phase, one-stage detectors like the You Only Look Once (YOLO) [6] series are favored over traditional two-stage methods due to their exceptional speed and efficiency, which are essential for real-time applications. This study utilizes the YOLOv10 [7] architecture, specifically chosen for its enhanced efficiency, architectural optimizations, and NMS-free inference capability, which are critical for low-latency localization of the small text regions containing container identifiers in dynamic port settings.

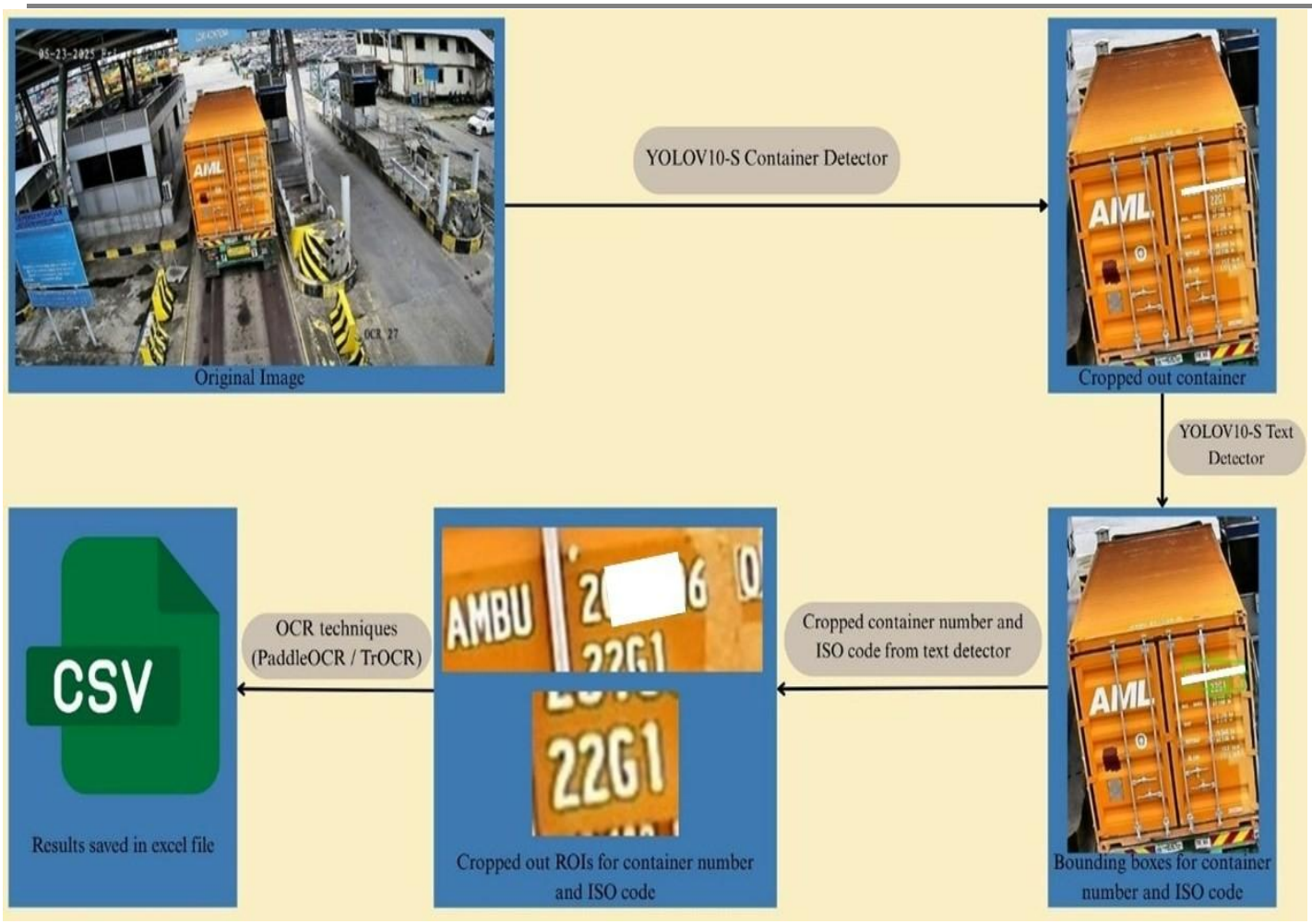

Figure 1: Overview process of container and ISO number recognition system with surveillance camera

For the subsequent recognition task, current state-of-the-art OCR models primarily diverge between speedoptimized and accuracy-optimized architectures. PaddleOCR [8] represents the speed-optimized approach, utilizing a lightweight Single Visual model for scene Text Recognition [9] (SVTR), often integrated with the EAST detector, making it highly suitable for applications demanding fast processing and high throughput. Conversely, TrOCR [10] (Transformer-based OCR) leverages complex Vision Transformer (ViT) architecture and self-attention mechanisms, excelling in accuracy by modeling contextual dependencies. This robustness makes TrOCR highly effective in handling highly distorted, faded, or complex text layouts common in challenging real-world container scenarios. Previous research, however, generally lacks a direct, rigorous, and systematic comparative evaluation of these two distinct architectural priorities—PaddleOCR (speed) versus TrOCR (accuracy) - when both are integrated within a unified YOLOv10 detection framework and tested under authentic, challenging port conditions. The deployment decision for automated systems hinges on resolving the critical operational trade-off between maximizing recognition speed (FPS) and ensuring the highest possible recognition accuracy (Exact Match).

In this study, limitations persist in achieving the simultaneous requirements of high-speed and high-accuracy recognition of container identifiers in challenging, real-world port environments characterized by variable lighting, low image quality, and complex backgrounds. This work addresses the need for a systematic, comparative evaluation of the leading speed-optimized (PaddleOCR) and accuracy-optimized (TrOCR) deep learning models when integrated with a modern, high-performance object detection framework (YOLOv10) to identify the optimal solution for operational container automation systems. The objectives of this study are: to design and implement a container automation system using YOLO for text detection combined with two OCR techniques: PaddleOCR and TrOCR for text recognition; to evaluate and compare the performance of PaddleOCR and TrOCR in terms of accuracy and speed when integrated into a container automation system; and to optimize and troubleshoot the system to enhance OCR performance and identify the most effective solution for container automation.

METHODOLOGY

The methodology section outlines the experimental setup used to design, train, and evaluate a multi-stage automated container recognition system. This system integrated the YOLOv10 object detection framework with two advanced Optical Character Recognition (OCR) models, PaddleOCR and TrOCR, under real-world port conditions.

A. Data Collection and Preparation

The study utilized a comprehensive dataset comprising 3,407 source images collected from surveillance cameras in real port environments, designed to capture diverse container types, orientations, and environmental challenges (see Figure 2). Images were manually annotated using Roboflow, focusing on establishing bounding boxes for the regions of interest (ROIs), specifically container numbers and ISO codes.

Figure 2: Collected surveillance camera images for labelling and detector training

To enhance model robustness and generalization, a structured preprocessing and augmentation pipeline was implemented with PaddleOCR. First, an auto-orientation function corrected the alignment of container text due to varied camera perspectives. All images were then uniformly resized to a consistent resolution of pixels to maintain efficiency and uniform input dimensions for training.

Data augmentation was performed primarily using bounding box cropping, which eliminated unnecessary background noise and focused the model's training exclusively on the text regions. This augmentation process expanded the initial dataset of 3,407 images to a final size of 8,899 augmented images. The total dataset was split into a training set (81%, 2,746 source images before augmentation) and a validation set (19%, 661 source images).

B. Experimental Design and Implementation

The container automation system was constructed as a two-stage pipeline:

1. **Container Detection (YOLOv10-S):** The YOLOv10 model was trained on the augmented dataset to serve as both a container detector and a text detector, localizing the container and precisely cropping the ROIs containing container numbers and ISO codes. The model was trained with 416 x 416 pixels input resolution, using cross-entropy loss to optimize both classification and localization accuracy.

2. **Text detection and recognition on the identified container cropped image with either:**

a. **PaddleOCR:** The PaddleOCR framework was fine-tuned using the cropped ROIs (text images) generated by the YOLO detector. Training utilized the SimpleData format and leveraged **Connectionist Temporal Classification (CTC) loss** to handle distorted and irregularly spaced text. Two distinct model configurations were tested: Config 1 prioritized accuracy, using the SVTR architecture augmented with a Spatial Transformer Network (STN) and 64 x 256 input pixel size, while Config 2 prioritized speed by omitting 48 x 256 pixels. Training was specialized, with separate models developed for container numbers and ISO codes.

b. **TrOCR:** The pre-trained, transformer-based TrOCR model was fine-tuned using the same set of cropped ROIs. The training was highly specialized to handle extreme real-world challenges (e.g., rotation, faded text), dividing the dataset by text type (ISO code vs. container number) and orientation (vertical vs. horizontal) to maximize recognition robustness.

C. Evaluation and Statistical Analysis

The performance of the integrated system was rigorously evaluated using a dedicated, unseen test set consisting of **173 images** sourced from an actual site deployment at the port terminal gate in Malaysia.

Detection Metrics (YOLOv10)

The performance of the detection stage was quantified using standard object detection metrics:

- **Mean Average Precision (mAP):** The primary metric measuring overall detection and classification accuracy across all text categories.

- **Intersection over Union (IoU):** Assessed the spatial accuracy of the predicted bounding boxes relative to the ground truth.

- **Inference Time:** Measured the processing speed of the YOLO detector (in seconds per image) to ensure suitability for real-time operation.

Recognition Metrics (PaddleOCR and TrOCR)

The accuracy and efficiency of the OCR models were assessed using specific text recognition metrics:

- **Exact Match Accuracy:** Measured the percentage of predictions that perfectly matched the ground truth text string, critical for container identification where small deviations result in functional errors (see Equation (1)).

$$Exact\ Match\ Accuracy = \frac{Number\ of\ Exact\ Matches}{Total\ Number\ of\ Predictions} \times 100 \quad (1)$$

- **Levenshtein Accuracy:** Based on the Levenshtein distance, this metric measured the minimum number of single-character edits (insertions, deletions, or substitutions) required to match the prediction to the ground

truth (see Equation (2)). This provides a more forgiving measure of how closely the prediction resembles the correct output.

$$Levenshtein\ Accuracy = 1 - \frac{Levenshtein\ Distance\ (Predicted, Ground\ Truth)}{Length\ of\ Ground\ Truth} \times 100 \quad (2)$$

- **Speed (Frames Per Second, FPS):** Measured the inference speed of each OCR model to quantify processing throughput for real-time application feasibility.

RESULTS AND DISCUSSION

This section presents the comparative performance results of the automated container recognition pipeline, analysing the trade-off between recognition accuracy and processing speed when utilizing the YOLOv10 detector paired with either PaddleOCR or TrOCR.

A. Performance of YOLOv10 for Container and Text Detection

The efficacy of the subsequent Optical Character Recognition (OCR) models relies fundamentally on the accurate and efficient localization of text regions by the object detection layer. The YOLOv10 model, selected for its single-pass efficiency over traditional two-stage detectors like Faster R-CNN, was evaluated for its localization precision.

On the full validation set (624 images), the model achieved a high Mean Average Precision (**mAP**) of **94.7%** and an average Intersection over Union (**IoU**) of **0.87**. This high IoU score confirms the spatial accuracy of the predicted bounding boxes relative to the ground truth.

Testing on the unseen, real-world deployment dataset (173 images from the port terminal gate), as shown in Table 1, confirmed robust operational readiness:

- **Container Detection Accuracy: 96.5%**.
- **Text Detection Accuracy (ROI localization): 97.6%**.

Achieving a text detection accuracy of 97.6% is critical, as it ensures that the high-quality Regions of Interest (ROIs) are provided to the OCR stage, minimizing errors caused by poor cropping or inclusion of background noise. This performance validates the use of a modern one-stage detector like YOLOv10, which provides the necessary speed and precision required for low-latency, real-time container automation environments.

Table 1: Performance Evaluation of Container and Text Detection Using Real-World Images from actual site deployment at the port terminal gate

Test Condition	Total Images	Correct Containers Detected	YOLO Container Detection Accuracy	Correct Text Detection	Text Detection Accuracy
Initial Test (Without YOLO Text Detection)	173	142	82.08%	121	70.8%
After Manual Cropping	50	N/A	N/A	44	88%
After Manual Cropping + YOLO Text Detection	50	N/A	N/A	49	97.66%

Final Test (Using YOLO Text Detection)	173	167	96.5%	163	97.6%

B. Comparative Recognition Accuracy of OCR Models

After detection, the cropped text regions were assessed for recognition accuracy using two primary metrics: Exact Match Accuracy (where the entire string must be perfectly correct) and Average Character Accuracy (which measures character-level correctness). TrOCR, the transformer-based model, consistently demonstrated superior recognition fidelity compared to PaddleOCR (see Table 2).

1. ISO Code Recognition

For ISO codes, which adhere to a standardized format, both models performed highly, but TrOCR maintained a decisive edge:

- **TrOCR Exact Match Accuracy: 98.73%**, with an Average Character Accuracy of **99.66%**.
- **PaddleOCR Exact Match Accuracy: 97.42%**, with an Average Character Accuracy of **99.52%**.

The robust performance of TrOCR in ISO code recognition is attributed to its transformer architecture, which leverages complex attention mechanisms to handle subtle distortions and complex layouts inherent in real-world port images. This capability ensures higher precision even when text is faded or rotated.

2. Container Number Recognition

Container number recognition is inherently more challenging due to the visual similarity between alphanumeric characters (e.g., "A" vs "M," "0" vs "O"), which increases the risk of misidentification in noisy environments. Consequently, the Exact Match Accuracy scores for this task were lower for both models:

- **TrOCR Exact Match Accuracy: 71.17%**.
- **PaddleOCR Exact Match Accuracy: 70.14%**.

TrOCR outperformed PaddleOCR by a margin of **1.03%** in inference accuracy for container numbers, confirming the value of its advanced architecture in ambiguous scenarios. This is crucial because errors like swapping an "A" for an "M" (e.g., predicting "MNBU" instead of "AMBU") are common but minimized by TrOCR's ability to better capture contextual relationships between characters.

Table 2: Accuracy Comparison of PaddleOCR and TrOCR

OCR Model	Exact Match Accuracy	Average Character Accuracy
PaddleOCR (Container Number)	70.14%	93.50%
PaddleOCR (ISO Code)	97.42%	99.52%
TrOCR (Container Number)	71.17%	94.13%
TrOCR (ISO Code)	98.73%	99.66%

C. Processing Speed Evaluation

Operational efficiency in container logistics depends heavily on throughput of the automated container text detection and recognition system, which is measured by processing speed (FPS). The experimental results, as

presented in Table 3, running on Intel Xeon CPU E5-2620v3, revealed a stark contrast, confirming the fundamental trade-off between model complexity (accuracy) and speed (FPS).

- **PaddleOCR Speed:** PaddleOCR, engineered for rapid processing via its lightweight architecture, achieved **18.35 FPS** for ISO codes and **10.90 FPS** for container numbers.

- **TrOCR Speed:** The TrOCR model, burdened by its larger, computationally demanding transformer architecture, was significantly slower, registering **7.93 FPS** for ISO codes and just **3.55 FPS** for container numbers.

PaddleOCR's high FPS makes it highly suitable for applications where containers arrive in rapid succession, as it processes images quickly enough to avoid operational bottlenecks. Conversely, TrOCR's lower speed is the necessary cost for achieving its higher recognition accuracy.

Table 3: Speed Comparison of PaddleOCR and TrOCR (Intel Xeon CPU E5-2620v3)

OCR Model	**FPS**
PaddleOCR (ISO Code)	18.35
PaddleOCR (Container Number)	10.90
TrOCR (ISO Code)	7.93
TrOCR (Container Number)	3.55

D. DISCUSSION

Optimizing the Accuracy vs. Speed Trade-Off

The comprehensive evaluation, as presented in Table 4, highlights that the selection of the optimal OCR model must be driven by the specific operational requirements of the port terminal.

The **YOLOv10 + TrOCR pipeline** provides the best overall performance when high precision is non-**negotiable**. Its accuracy advantage is crucial for minimizing the risk of costly misclassification errors, especially when dealing with degraded, worn, or rotated container labels. Although slower (3.55 FPS for container numbers), the improved fidelity in text extraction mitigates financial and logistical risks associated with incorrect tracking. The integrated system achieved an overall recognition accuracy of **92%** on the test dataset.
In contrast, the **YOLOv10 + PaddleOCR pipeline** is ideal for high-throughput, **real-time applications** where maximizing container flow and speed is prioritized over maximum theoretical accuracy. PaddleOCR's speed (up to 18.35 FPS) ensures that the automation system can rapidly process large volumes of images, which is essential for busy ports where operational lag must be avoided.

The results emphasize that architectural choice dictates performance: the complexity of TrOCR's transformer design yields superior accuracy in challenging text recognition, while the efficiency of PaddleOCR's lightweight design delivers the requisite speed for critical real-time processing.

Table Error! No text of specified style in document.**:** OCR Model Comparison by Speed, Accuracy, and Use Case

Model	**Speed (FPS)**	**Accuracy (Exact Match)**	**Ideal Use Case**
PaddleOCR	Fast (10.90 FPS)	Lower (70.14% for container numbers)	Real-time applications where speed is essential

TrOCR	Slow (3.55 FPS)	High (71.17% for container numbers)	Applications requiring high accuracy especially with distorted text.

CONCLUSION

This research successfully designed, implemented, and critically evaluated a robust, two-stage container text recognition system aimed at optimizing automated logistics processes in challenging port environments. The primary research question concerning the optimal balance between accuracy and speed was answered by comparing the performance of the lightweight PaddleOCR and the complex TrOCR models when integrated with the efficient YOLOv10 detector. The study concludes that the YOLOv10 with TrOCR pipeline provides the best overall solution for reliable, high-accuracy container and ISO code recognition, achieving a system-wide recognition accuracy of 92% on the real-world test dataset. Specifically, TrOCR demonstrated superior recognition fidelity, achieving 98.73% exact match accuracy for ISO codes and 71.17% for container numbers, slightly outperforming PaddleOCR in critical accuracy scenarios. While PaddleOCR achieved significantly faster throughput (up to 18.35 FPS for ISO codes), TrOCR's transformer-based architecture proved more effective in mitigating errors caused by distorted, faded, or visually ambiguous characters, yielding a 0.95% increase in container number accuracy.

The practical implication of these findings offers clear guidance for port automation deployment: terminals where precision is paramount (e.g., minimizing costly errors in container routing or customs documentation due to misclassification) should favour the YOLOv10 with TrOCR pipeline. This robust solution ensures high confidence in data extraction despite unfavourable environmental conditions. Conversely, the faster PaddleOCR is better suited for high-throughput, real-time applications where maximum speed is necessary to prevent operational bottlenecks in high-volume processing environments. This work contributes to advancing automated port logistics by providing an effective, scalable, AI-powered system that successfully balances efficiency with the stringent accuracy requirements of container identification.

Despite its robust performance, the system has certain limitations. The project was strictly confined to the OCR component and did not include integration or evaluation with other logistical elements of a complete container automation system. Furthermore, although substantial data augmentation was utilized, the model's overall accuracy may still be constrained by the diversity of the publicly available image datasets used for training, which may not fully represent all extreme real-world variations encountered in complex operational environments.

Based on these findings, future work should focus on integrating a compact lexicon library, specifically tailored to frequently occurring ISO codes and container text sequences, to guide the OCR models. This specialized lexicon would enhance both recognition speed and accuracy by constraining the prediction space, helping the models handle unusual or rare identifier combinations more efficiently.

ACKNOWLEDGEMENT

The authors would like to express thanks to collaborator at port terminal and Faculty of Electronics and Computer Technology and Engineering (FTKEK) at Universiti Teknikal Malaysia Melaka (UTeM) for their assistance in acquiring the essential information and resources for the successful completion of the research.

REFERENCES

1. Port of Rotterdam. Sustainability and Innovation in the World's Most Automated Port. Retrieved from https://www.portofrotterdam.com/en/ourport/facts-figures/port-innovation (accessed 9 March 2025).
2. Yang, J. Advanced Automation at China's Major Ports: A Study of Shanghai and Qingdao. Journal of Transport and Logistics, 2023. Retrieved from https://www.journals.elsevier.com/journal-of-transportand-logistics (accessed 9 March 2025).

3. Australian Maritime Safety Authority. Enhancing Container Tracking with OCR Systems in Australian Ports. Infrastructure Review, 2021. Retrieved from https://www.amsa.gov.au/safety-navigation. (accessed 9 March 2025).
4. Australian Maritime Safety Authority. OCR Implementation for Container Recognition at Australian Ports. Maritime Technology Reports, 2021. Retrieved from https://www.amsa.gov.au/marine-technology. (accessed 10 March 2025)
5. S. Jie, "The world's largest single fully automated terminal Shnaghai Yangshan Port Phase IV opening," China Economic Weekly, pp. 58–60, 2017.
6. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Computer Vision and Pattern Recognition (CVPR), 2016.
7. A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, and J. Han, "Yolov10: Real-time end-to-end object detection," Advances in Neural Information Processing Systems, vol. 37, pp. 107984–108011, 2024.
8. C. Li et al., "PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system," arXiv preprint arXiv:2206.03001, 2022.
9. Y. Du et al., "SVTR: Scene text recognition with a single visual model," in Proc. Int. Joint Conf. Artif. Intell. (IJCAI), 2022, pp. 888–896.
10. M. Li et al., "TrOCR: Transformer-based optical character recognition with pre-trained models," in Proc.
11. AAAI Conf. Artif. Intell., vol. 37, no. 11, pp. 13094–13102, Jun. 2023.