# A Review of Text Recognition in Complex Scene Images Based on Deep Learning

Xuchen Wang

Country College of Science, Southwest Petroleum University, 637001, Nanchong, China
202331095107@stu.swpu.edu.cn

**Abstract.** With the rapid development of computer vision technology, Scene Text Recognition (STR) in complex scenarios has become a critical technology in intelligent security, autonomous driving, augmented reality, and related fields. However, traditional Optical Character Recognition (OCR) methods show inherent limitations in complex environments, including poor scene adaptability, reliance on manual feature extraction, and weak anti-interference capabilities. This paper systematically reviews advancements in deep learning-based STR technologies for complex scenarios, analyzing the entire workflow from text detection/recognition to end-to-end systems. Through practical applications such as license plate recognition, tunnel defect detection, and object sorting, we validate the effectiveness of YOLO and CRNN models in specific scenarios, while identifying bottlenecks including excessive model parameters and insufficient hardware compatibility. Further analysis indicates persistent challenges: high computational resource consumption, limited model generalization capabilities, poor interpretability, inadequate real-time performance, and vulnerability to adversarial samples. Finally, this paper proposes future research directions focusing on: Multimodal feature fusion, Weakly supervised learning paradigms, Lightweight model deployment and Decision-centric learning mechanisms These approaches aim to promote practical STR implementations in complex scenarios through interdisciplinary technological collaboration.

**Keywords:** Computer vision, deep learning, scene text recognition technology, Complex scenarios.

## 1 Introduction

Text and image recognition technology has been widely adopted across diverse domains and is closely intertwined with daily life. However, such as adverse weather conditions like rain and fog can degrade transmission media, leading to reduced contrast and significant challenges in target identification.

Conventional computer systems often struggle to directly detect and recognize real-world images and text. Traditional algorithms for complex scene image recognition face limitations in scene adaptability, heavy reliance on manual feature extraction, and susceptibility to localization/recognition errors. Current approaches predominantly


J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_75

integrate existing technical frameworks with artificial intelligence (AI) to implement deep analytical recognition [1].

Researchers mainly use OpenCV and Yolo models to design programs. In the image preprocessing stage, Gaussian filtering is first applied for noise reduction to enhance the accuracy and stability of the license plate recognition system. Secondly, color space conversion is performed to transform images from color to grayscale, reducing color interference. Sobel operators, Canny operators, and Laplacian operators are applied for license plate localization. The training of convolutional models utilizes license plate datasets (containing numerous images with Chinese provincial characters, numbers, and English letters) to construct and train convolutional neural networks, ultimately employing neural network models for license plate recognition.

For different application scenarios, minor modifications to existing models can achieve desired objectives. For example, when adopting CRNN models, researchers replaced 8 traditional convolutional layers in VGG-16 with generalized shuffle convolution (GSConv) to avoid the massive network parameters and model size caused by deep network structures (which increase deployment costs and implementation difficulties when deploying to mobile devices). This modification reduces the computational load in convolutional layers while achieving model compression and improved operational speed [2][3].

Scene Text Recognition (STR) in Complex Environments: As an important research direction in computer vision, STR demonstrates significant value in applications such as intelligent security, autonomous driving, and augmented reality. This paper conducts a horizontal analysis of evaluation metrics and performance comparisons in current mainstream datasets (e.g., ICDAR, COCO-Text), systematically reviewing the development of deep learning-based STR technologies in complex scenes.

The study specifically analyzes critical technological breakthroughs and limitations in preprocessing text detection, text recognition, and end-to-end systems under complex scenarios, proposing that future research should focus on multimodal feature fusion, weakly supervised learning paradigms, and lightweight deployment. The technical evolution from convolutional neural networks (CNN), recurrent neural networks (RNN) to attention mechanisms and Transformer architectures is thoroughly discussed.

For challenging scenarios involving multi-directional text, curved text, and low resolution, this paper provides in-depth exploration of instance segmentation-based detection methods.

Firstly, we reveal solutions for multi-oriented text, curved text, and low-resolution scenarios. by examining the technical evolution of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), attention mechanisms, and Transformer architectures, Comparative analysis demonstrates the performance advantages of instance segmentation-based detection methods on mainstream datasets (e.g., ICDAR, COCO-Text).

Secondly, through practical applications such as license plate recognition, tunnel defect detection, and object sorting, we validate the effectiveness of YOLO and CRNN models in specific scenarios, while identifying bottlenecks including excessive model parameters and insufficient hardware compatibility.

Further analysis indicates persistent challenges: high computational resource consumption, limited model generalization capabilities, poor interpretability, inadequate real-time performance, and vulnerability to adversarial samples.

Finally, this paper proposes future research directions focusing on: Multimodal feature fusion, Weakly supervised learning paradigms, Lightweight model deployment and Decision-centric learning mechanisms.

## 2 Convolutional Neural Networks (CNN) for Deep Learning

Convolutional Neural Networks possess excellent learning capabilities and information storage capacity. The CNN is composed of convolutional layers, pooling layers, activation functions, and fully connected layers. The convolutional layer extracts local features and spatial relationships through convolution operations on image patches. The pooling layer reduces the size of feature maps to decrease computational load and prevent overfitting while extracting more prominent features. Three pooling methods are available: global pooling, average pooling, and max pooling. Max pooling is typically used because it preserves feature map information to the greatest extent. Activation functions fit complex nonlinear data and activate neurons, enabling neural networks to model intricate patterns. The fully connected layer performs classification and regression tasks.

### 2.1 CRNN for deep learning

The CRNN model divides images into segments, with each segment undergoing character prediction. The CRNN architecture consists of three modules: a feature extraction module, a sequence modeling module, and a text decoding module. During feature extraction, a deep convolutional neural network processes input images (typically resized to a height of 32 pixels), extracting hierarchical visual features through multiple convolutional layers.

After feature extraction, the spatial width of the output feature map is reduced to one-fourth of the original, and the channel dimension expands to 512. The recurrent network layer uses a bidirectional LSTM (Bidirectional Long Short-Term Memory Network) to capture forward and backward correlations in feature sequences.

The CRNN network includes 7 convolutional layers and 4 max-pooling layers. This structure has a large parameter count, high memory consumption, and significant training difficulty.

### 2.2 OpenCV for image processing

OpenCV holds significant advantages in image processing. For example, applying the Sobel operator enhances image features, effectively improving robustness in complex scenes.

### 2.3 YOLO for image processing

YOLO ("You Only Look Once") transforms object detection into a regression problem, completing detection in a single forward pass without candidate region generation and filtering (e.g., YOLOv8 achieves hundreds of frames per second).

YOLO processes the entire image in a single inference, capturing global contextual information and reducing false detections caused by partial occlusion or background interference. For instance, it infers partially obscured vehicles or pedestrians through environmental context.

## 3 Current Research Examples

### 3.1 Tunnel defect detection

Researchers improved feature extraction by adjusting network weights multiple times. YOLOv3-based recognition technology enhances both speed and accuracy [4-7]. To address the precise detection of tunnel leakage and rock spalling, a customized dataset was constructed, and an intelligent recognition model was developed based on the YOLOv3 framework. The detection system integrates a deep convolutional feature extraction module and a multi-task prediction module, utilizing detection units at 13×13, 26×26, and 52×52 scales for multi-scale target analysis. Each unit employs optimized anchor parameters for adaptive feature matching, with iterative parameter refinement to improve recognition accuracy.

### 3.2 Object sorting

Researchers compared custom models with AlexNet, VGG16, and GoogLeNet [8]. They found out that Validation accuracy can be 99.6%~100 %. And the Validation loss is 0.001~0.0118. The Training-validation accuracy gap: 0~0.003%, indicating minimal overfitting.

All models demonstrated high recognition performance, with excellent loss fitting on the validation set and outstanding results on fruit datasets. AlexNet, VGG16, and GoogLeNet achieved training accuracies of 99.66%, 99.65%, and 99.9%, and validation accuracies of 96.5%, 99.9%, and 99.6%, respectively. These models meet practical requirements for real-world fruit image recognition.

### 3.3 Agriculture

Current maturity detection for agricultural products uses convolutional neural networks (CNNs) and artificial neural networks (ANNs) to extract color and volume features. Deep CNNs achieve high accuracy and efficiency in maturity detection. Challenges include high costs due to large target numbers and difficult image acquisition, as well as limited superiority over manual methods.

Applications are mostly limited to pre-harvest stages, with few studies incorporating texture and shape features. A proposed image recognition system compares color and volume features with mature bananas, achieving 97.75% accuracy. Kheiralipour et al. classified wild pistachios into four maturity stages using linear discriminant analysis (93.75%), quadratic discriminant analysis (97.50%), and ANNs (100% accuracy) [9-11].

## 4 Defects and Limitations of Deep Learning-Based Image Recognition in Complex Environments

Training deep neural networks (e.g., ResNet, YOLO) requires high-performance GPU clusters and distributed storage systems, resulting in substantial computational resource consumption and hardware costs that are prohibitive for individual developers or small-to-medium enterprises [12].

Model Compression Challenges: Although techniques like pruning, quantization, and lightweight networks (e.g., MobileNet) have been proposed, reducing computational complexity while maintaining accuracy remains technically demanding. Model compression and optimization continue to pose significant challenges.

Insufficient Model Generalization: Issues include overfitting and poor scene adaptability. Current technologies struggle to ensure safety and reliability in complex environments. While accuracy often exceeds 90%, it remains inadequate for high-risk scenarios (e.g., autonomous driving systems failing in rain/snow).

Cross-domain migration is difficult due to varying feature requirements across disciplines, necessitating model retraining and redesign.

Poor Interpretability: The black-box decision-making mechanism of deep learning models lacks transparency in feature extraction and classification, raising ethical and practical risks in critical applications like medical diagnosis. Visualization techniques (e.g., attention mechanisms, feature maps) partially reveal model focus areas but fail to reconstruct full decision logic.

Real-Time and Deployment Efficiency: Complex models (e.g., Faster R-CNN) face inference latency on edge devices (e.g., drones, mobile terminals), hindering practical deployment. Hardware compatibility issues arise when optimizing algorithms for diverse architectures (e.g., FPGA, ASIC).

Vulnerability to Adversarial Attacks: Deep learning models (e.g., neural networks) exhibit linear responses to input perturbations in high-dimensional spaces. Adversarial samples crafted via gradient-based methods (e.g., FGSM) can cause misclassification with minimal pixel-level modifications.

# 5 Future Prospects for Deep Learning-Based Image Recognition

## 5.1 Synergistic mechanisms between deep neural networks and feature extraction

The integration of deep neural networks and intelligent feature extraction is reshaping computer vision. Key advancements include:

Deep Feature Modeling, Convolutional neural networks enable hierarchical visual feature modeling through multilayered nonlinear mappings. End-to-end training improves feature recognition accuracy by ~35% and enhances noise resistance.

Feature Optimization: Intelligent feature selection reduces data redundancy and accelerates model convergence by ~40%. Attention-based visualization techniques partially decode black-box decision logic, enhancing interpretability.

## 5.2 Evolution of decision-centric learning in vision

Decision-centric learning frameworks are revolutionizing complex visual tasks:

Image Enhancement means in super-resolution tasks, dynamic policy adjustments achieve multi-scale feature fusion, improving peak signal-to-noise ratio (PSNR) by 2.8 dB compared to traditional methods. Adaptive noise suppression reduces the mean squared error to 0.032 while preserving details.

Intelligent Perception and Path Planning means Decision-driven systems achieve centimeter-level localization accuracy in navigation. Multimodal frameworks integrating visual, depth, and motion data boost obstacle avoidance success rates to 97.6%.

Meta-learning-based path planning reduces new-scene adaptation time by 80%, enhancing environmental adaptability for autonomous driving and robotics.

# 6 Conclusion

As a frontier direction in computer vision, scene text recognition (STR) technology in complex scenarios has achieved significant progress driven by deep learning. Through a systematic review of STR technology's evolutionary trajectory, this paper validates the core roles of CNN, RNN, attention mechanisms, and Transformer architectures in text detection/recognition tasks. Specifically， YOLO-based end-to-end detection frameworks dramatically improve real-time performance CRNN models effectively solve curved text recognition challenges through sequence modelling Lightweight improvement strategies (e.g., generalized shuffle convolution) to alleviate resource constraints in model deployment.

Application cases including license plate recognition, tunnel defect detection, and agricultural maturity assessment demonstrate that deep learning technologies can

significantly enhance robustness and recognition accuracy in complex scenarios, with accuracy exceeding 99% in specific cases.

However, current technologies still face multiple challenges: First, deep model training relies on high-performance hardware, with prohibitive computational costs limiting adoption in small/medium-scale applications.

Second, Insufficient model generalization causes cross-domain migration difficulties, particularly showing performance plummets under rain/snow or extreme illumination;

Third, Black-box decision mechanisms undermine system credibility, while high-risk scenarios like healthcare/autonomous driving urgently demand interpretability;

Fourth, Adversarial sample attacks expose model fragility, where minor pixel perturbations may trigger critical misjudgements;

Fifth, Edge device inference latency and cross-platform compatibility issues still require optimization.

Future research directions should prioritize addressing computational complexity through Neural Architecture Search (NAS), dynamic pruning, and quantization while preserving recognition accuracy.

Concurrent efforts could establish cross-modal understanding frameworks to enhance scene adaptability and enable effective fusion of multimodal features.

Furthermore, developing attention mechanisms with complementary feature visualization techniques could significantly improve model interpretability.

A critical challenge lies in establishing comprehensive adversarial defense mechanisms to strengthen immunity against environmental noise and malicious perturbations.

Additionally, implementing meta-learning and reinforcement learning frameworks may facilitate cross-domain knowledge transfer and enable rapid adaptation to novel scenarios.

With advancements in hardware capabilities and algorithmic innovations, scene text recognition (STR) technology is poised to achieve broader deployment in smart city infrastructures, industrial quality inspection systems, and medical diagnostic workflows, ultimately facilitating the transition of AI systems from laboratory environments to real-world applications that address practical challenges.

## References

1. Zhang, K.: ‘Design of license plate recognition system based on OpenCV’, Integr. Circuit Appl., **40**(7), pp. 198–199 (2023)
2. Rao, Z.: ‘Research on license plate recognition system in unstructured scenes based on deep learning’, Master thesis, Hunan Univ., Changsha, China, pp. 1–85 (2022)
3. Zhang, Z., Fu, R., Tian, Q.: ‘Research on target detection method of gated image based on YOLO-v8’, J. North China Univ. Technol., 23(3), pp. 45–52 (2024)
4. Lu, Z.: ‘Research on rapid detection method for tunnel defects based on deep learning’, Inf. Build. Constr., **42**(2), pp. 112–118 (2025)
5. Zhu, D., Cheng, X., Qiu, Y., et al.: ‘Advances in crop image recognition technology based on deep learning’, J. Agric. Mech. Res., **46**(1), pp. 20–25 (2025)

6. Lin, L., Zhuang, Y., He, H., et al.: 'Improved CRNN-based license plate recognition method', J. Meas. Sci. Instrum., **15**(6), pp. 78–84 (2024)
7. Li, C., Guo, X., Yang, J., et al.: 'Fruit image recognition based on deep learning', J. Chin. Agric. Mech., **46**(1), pp. 198–203 (2025)
8. Cai, R., Huang, Z., Peng, J.: 'Pear sorting system based on convolutional neural network', J. Sens. Technol., **37**(11), pp. 33–39 (2024)
9. Li, Z.: 'Research on license plate localization algorithm in complex scenarios based on improved YOLO model', Master thesis, Southwestern Univ. Finance Econ., Chengdu, China, pp. 1–76 (2023)
10. Ren, R., Wang, X., Wen, C.: 'Improved CRNN-based Chinese street view text recognition technology', J. Chengdu Univ. Inf. Technol., **28**(2), pp. 55–62 (2025)
11. Lee, J., et al.: 'Identification and classification of images in e-cigarette-related content on TikTok', Subst. Use Misuse, **60**(5), pp. 677–683 (2025)
12. Camargo, C., Fernández, A.: 'Neuropedagogy and neuroimaging of artificial intelligence and deep learning', Educ. Process Int. J., **13**(3), pp. 97–115 (2024)