2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Distributed Hierarchical Structure for Multi-Objective OCR Text Recognition in Electrical Cabinet Inspection

Zhixin Kang[1] , Shu Lin[2] , Jungang Xu[3] and Qingjie Kong[4]

[1,2,3] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China

kang-zhixin22@mails.ucas.ac.cn;slin@ucas.ac.cn;xujg@ucas.ac.cn
[4] Riseye Robotics (Shanghai) Co., Ltd., Shanghai 201100, China
qjkong@riseye.ai

Abstract. Electrical Cabinet Label Recognition is an important part of robotic intelligent inspection. Accurate label recognition is a prerequisite for effectively recording inspection anomalies. After the inspection robot takes pictures, OCR (Optimal Character Recognition) can detect the location of the electrical cabinet labels and recognize the content on the labels. Firstly, We use the oclip method to reduce the model training time and reduce the need for dataset size. Secondly, In the electrical cabinet label dataset, the text location detection accuracy based on the cutting-edge model DBNet++ reaches 96.74%, and the text content recognition accuracy based on ABINet reaches 89.33%. Through comparative experiments, we found that applying only the ABINet visual model can improve the text recognition accuracy to 90.58%, indicating that the language model in ABINet does not perform well for this task. The ABINet visual model is better at extracting local information from the image text, while the ABINet language model excels at recognizing semantic relationships across different parts of the text. Thirdly, Leveraging this characteristic, we designed a distributed hierarchical structure for the multi-objective OCR text recognition framework ABINet-TS. In the first layer, the visual model is used to recognize local information from the image, while in the second layer, the language model is applied to correlate and correct the predictions made by the visual model. This further improves the text recognition accuracy to 91.74%.We further replace the language model in ABINet-TS with BERT, which further improved the accuracy of text lines to 92.16%.

Keywords: OCR, Text recognition, Text detection, Deep learning.

1 Introduction

Substation inspection is one of the daily operation and maintenance tasks at substations. By inspecting and checking the status of high-voltage and low-voltage cabinets, faults or hidden dangers in the electrical cabinets can be identified immediately, allowing for timely repairs and preventive measures, ensuring the safety and stability of the substation. In actual operations, substation inspections are often carried out manually. However, manual inspections have many drawbacks: long inspection cycles with extended periods of interruption; the possibility of missed or incorrect inspections; and high labor costs.

Robot Intelligent Inspection is a process where a wheeled robot takes pictures along a fixed route within a substation. Using hardware and software technologies, the robot performs recognition and detection of the status of electrical cabinets in the substation. The robot is equipped with optical cameras, infrared cameras, and other devices to capture images of the electrical cabinets. Using target detection models like YOLOv7 [1], it is possible to quickly and efficiently detect the status of components on the electrical cabinets, such as the color of indicator lights and the orientation of knobs. Based on the status of the various components on the electrical cabinet, it can be determined whether the cabinet being photographed has any abnormalities. Through OCR models, the location of abnormal electrical cabinets can be quickly pinpointed, and the labels of the abnormal cabinets can be recognized. This enables the reading, matching, and recording of the abnormal cabinet and its label, allowing faults to be detected promptly. Compared to traditional manual inspections, intelligent inspections significantly enhance the safety and reliability of the inspection process and drastically reduce the labor costs associated with inspections.

Substation Electrical Cabinet Label Detection which can quickly locate abnormal labels is a crucial part of intelligent inspection. After the inspection robot captures images, the OCR text detection model detects the position of the electrical cabinet label in the image, and the OCR text recognition model predicts the text characters based on the image of the text. This inspection method enables quick, automatic, and precise localization of abnormal electrical cabinets. By using the coordinates of the inspection robot during image capture and the label on the captured image, a preliminary electrical cabinet label database for the substation can be constructed, reducing the cost of manually building the database. The accuracy and speed of OCR recognition have continuously improved in recent years. OCR model frameworks such as PaddleOCR and TesseractOCR can efficiently and accurately detect and recognize text in images across many application scenarios.

Although OCR technology has made significant breakthroughs, directly applying existing frameworks in substation scenarios still does not yield satisfactory results. The accuracy requirements for electrical cabinet label recognition in substations cannot be met. We propose a distributed hierarchical structure for the multi-objective OCR text recognition framework ABINet-TS. This framework decomposes complex text recognition tasks, fully leverages the advantages of both visual models and language models , and effectively improves the accuracy of text recognition models for electrical cabinet label recognition. It efficiently and accurately meets the demands of robotic intelligent inspection. The experimental process for electrical cabinet label detection and recognition is shown in Figure 1.

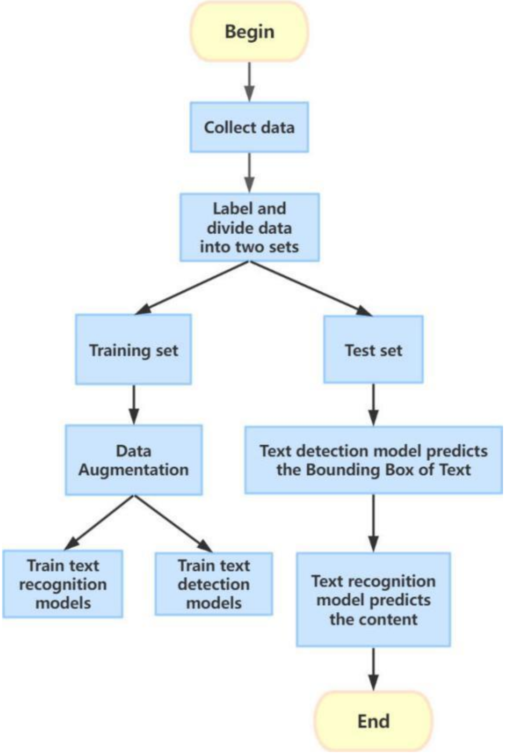
Begin
Collect data
Label and divide data into two sets
Training set
Test set
Data Augmentation
Text detection model predicts the Bounding Box of Text
Train text recognition models
Train text detection models
Text recognition model predicts the content
End

Fig. 1 Electrical Cabinet Label Detection Flowchart

2 Related works

OCR can be mainly divided into two models: the text detection model for locating the position of text boxes and the text recognition model for recognizing the content of text in images. The following will introduce the current research status of OCR from these two aspects.

2.1 OCR Text Detection Model

OCR text detection is the task of detecting and marking the location of the target text regions in an image. Traditional text detection methods usually rely on handcrafted features to detect character regions. Epshtein et al. [2] proposed the Stroke Width Transform (SWT) to detect characters. Neumann and Matas [3] proposed a method for locating characters by classifying Extremal Regions. They see the character detection problem as the process of selecting valid sequences from a set of extreme regions. The detected extreme regions are then combined into words. FASText [4] is a fast text detection system that applies the FAST key point detector to stroke detection. In recent years, deep learning-based methods have significantly outperformed traditional methods in terms of speed and adaptability. Especially in challenging scenarios, such as low resolution and geometric distortion, deep learning models can greatly improve recognition accuracy and are more robust to environmental changes. Deep learning-based text detection methods are mainly divided into two categories: regression-based text detection methods and segmentation-based text detection methods.

Regression-based methods

The TextBoxes method proposed by Liao et al.[5] modifies the anchor and scale of convolutional kernels to achieve text detection. TextBoxes++ [6] applies a quadrilateral regression method to detect text with multiple orientations. SSTD[8] applies attention mechanisms to roughly identify areas of text. EAST [9] proposes a two-stage text detection method: a fully convolutional neural network and Non-Maximum Suppression (NMS).

Segmentation-based methods

Segmentation-based methods typically first perform pixel-level predictions, then use some post-processing algorithms to combine the predicted pixels into predicted regions. These post-processing algorithms are often more complex and time-consuming. Zhang et al. [11] by semantic segmentation and based on the most stable extreme region. The algorithm of the domain detects text in multiple directions. Xue et al. [12] proposed text edges Text border semantics and a kind of bootstrapping Technical segmentation of text instances. PSENet [13] introduced generating kernels of different scales for each text instance and progressively expanding the smallest scale kernel to form the final complete text prediction. Due to the large gaps between the smallest scale kernels, this method can efficiently separate neighboring text instances, making segmentation-based methods more suitable for detecting text instances of arbitrary shapes. SegLink++[16] enables dense and arbitrarily shaped scene text detection with instance-aware components and minimal spanning trees. DBNet++ [15] proposed a differentiable binarization method that performs binarization operations in the segmentation network, allowing the network to adaptively adjust the binarization threshold. The proposed adaptive scale fusion module can reuse multi-scale feature maps adaptively.

2.2 OCR Text Recognition Model

OCR text recognition aims to identify the text content in an image, which is a task of inferring text from images. OCR text recognition can primarily be divided into two approaches: vision-based methods and vision-language model-based methods.

Vision-based models

Vision-based models typically consist of two parts: the visual model that extracts features from the image, and the translation method that converts the model's predicted characters into text. Shi et al. [17] extracted features using CNN and LSTM, and then used the CTC method [18] to obtain the predicted sequence. He et al. applied a deep text recurrent network [21], while Hu et al. proposed a CTC decoder guided by a graph convolutional network.[23]. Du et al. proposed an encoder similar to ViT [24]. Models based on visual frameworks do not fully consider the relationships between context in the text, resulting in lower recognition accuracy.

Vision-Language methods

Vision-Language based models [30] leverage semantic information from the text to correct the features extracted by the visual model. Lee et al. [31] were the first to introduce the attention module into the text recognition task. Cheng et al.[33] added some attention modules to allow the model to learn more features.

Cheng et al.[34] extracted text features in four directions. Wojna et al.[35] added positional coding of feature maps to enhance sequence order. Lyu et al.[36] replaced RNN with Transformer[32] to improve the recognition accuracy of the model. VisionLAN [37] introduced a word-by-word masking learning method to endow the visual model with language capabilities. During inference, only the visual model is applied, which improves the model's inference speed. Atienza et al. [39] proposed that ViT could be used as a feature extractor for visual models. SVTR [40] first decomposes text images into small blocks, and after blending these blocks, it uses both global and local mixed attention to capture patterns between and within characters.

3 Proposed methods

The model for recognizing text labels is based on the ABINet [41] framework. ABINet is the best-performing model among state-of-the-art models for recognizing images of electrical cabinet labels in substations. ABINet consists of two parts: a visual model and a language model. The vision model extracts feature from the image, then the language model corrects the predictions made by the visual model through a multi-head attention module with masking. Finally, the probability matrices predicted by the visual model and the language model are combined through weighted averaging to generate the final fused prediction result. The language model is autoregressive, meaning its output can be fed back as input multiple times to iteratively refine and correct the predictions. ABINet is very good at predicting images that contain natural language.

3.1 ABINet-V

ABINet-V retains the visual model part of ABINet but modifies the loss function to a cross-entropy loss function. The visual model part is composed of ResNet [42] and position attention module. ResNet extracts features from the image, the features extracted by the multi-layer ResNet are then combined, and the features extracted from different layers of ResNet are merged using a multi-scale fusion module. The positional attention module focuses on the positional information of the predicted features and further extract features. The linear layer then generates the predicted sequence.The visual model part of ABINet does not consider the semantic information of the text in the image, instead extracting features from the image and predicting the text sequence. In cases where the semantic correlation between text characters is low, the correction effect of the language model is weakened, making the visual model's prediction results particularly important. In the substation electrical cabinet label dataset, the text in the images contains limited semantic information, such as "4kW 8.2A". Therefore, on this dataset, ABINet-V is likely to perform better.

3.2 ABINet-TS

ABINet-TS (Two-Stage) is a two-stage model, as shown in Figure 2. First stage: ABINet-V simultaneously recognizes multiple individual text images, using ResNet and the Positional Attention module to extract features, and finally predicts the text sequences. Second stage: The model recognizes and corrects the contextual relationships of the visual model's predictions, predicting the sequence of associated text. In the first stage, ABINet-V is trained on a single-label dataset, while in the second stage, ABINet-L is trained on a concatenated image dataset. The two-stage model ensures that the visual model part of ABINet and the language model part do not interfere with each other. The visual model is responsible for predicting individual label images, ensuring high accuracy for predictions of single-label images. The language model is responsible for correcting the predictions of the visual model by considering the semantic relationships between the label texts.

The input to the language model comes entirely from the output of the visual model, so the accuracy of the visual model's predictions is crucial to the final model's output. The loss function of ABINet includes the loss functions of the visual model, the language model, and the fusion module. The concatenated image dataset contains more information, which makes it difficult for the visual model in ABINet to predict accurately on the concatenated image dataset. The visual model of ABINet-TS is first trained on a single-label dataset. The prediction results of the visual model do not need to consider the semantic context of the labels, ensuring high accuracy in recognizing individual labels. In ABINet-TS, the language model builds on the accurate predictions of individual labels from the visual model. It then incorporates the semantic relationships of contextual labels, enabling more accurate prediction of concatenated label text sequences. Furthermore, the iterative correction process of the language model also allows the model to more fully learn the contextual relationships between labels.

The language model and visual model in ABINet-TS are decoupled, meaning that different model combinations can be used for each, providing greater flexibility and generalization of the model. The performance of the language model depends on the input from the visual model. Therefore, using a visual model with better recognition performance will improve the recognition accuracy of ABINet-TS. Similarly, selecting a language model with stronger semantic analysis capabilities can help the model learn more about the contextual semantic relationships.

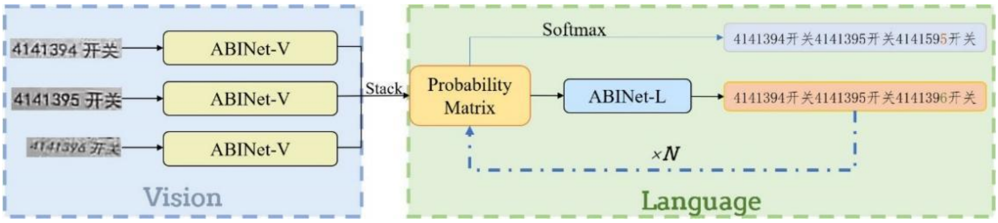

Fig. 2 Architecture of ABINet-TS

3.3 Oclip

Oclip[43] is a pre-training method for OCR model parameters. This approach reduces the time it takes to train the model and the dataset required for model training.

As shown in Fig 3, the visual encoder part of the OCLIP extracts features from the image and generates a sequence of image features. After the text sequence with some characters occluded is positionally encoded, the features are extracted by the oclip language encoder to generate a text feature sequence. The image feature sequence is input into the visual-text decoder as Q (Query), and the text feature sequence is input into the visual-text decoder as K (Key) and V (Value), and finally the predicted occluded characters are generated. The sum of losses is lost by the comparison of the text feature sequence with the text feature sequence. The classification loss of the predicted occluded character vs. the actually occluded character to update the model parameters. The parameters of some models can be migrated to other models for direct use.

The oclip method makes the backbone of the model pay more attention to the text information in the image, which allows the model to extract more. The features of the picture Chinese book are helpful for subsequent OCR model training and save the amount of data required for model training. Compared with other pre-training methods, the oclip pre-training strategy pays more attention to the correlation between images and text, and the OCR text recognition task is to recognize the text in the image from the image, so the oclip model parameter pre-training strategy is more suitable for the actual needs.

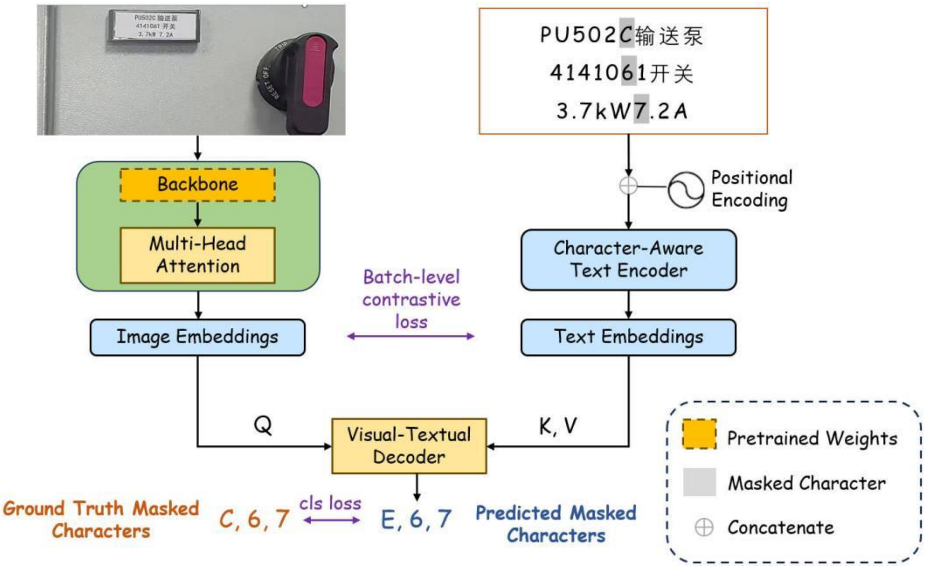

Fig. 3 Architecture of oclip

3.4 ABINet-TSB

The prediction results of the ABINet-TS model rely too much on the prediction results of the first tag in a series of tags, which may lead to the prediction results of subsequent texts being subject to the prediction results of the first word. In order to solve the problem of contextual association of label text, we introduce the BERT model. Implemented a "cloze in the blank" style of contextual text line prediction reference. We call the model in which the language model in ABINet-TS is replaced by the BERT method as ABINet-TSB.

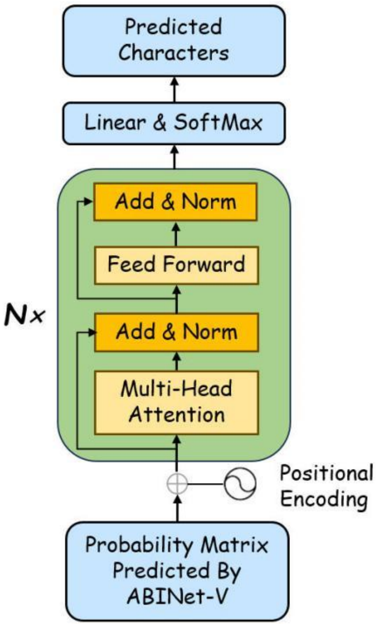

Fig. 4 Architecture of BERT

As shown in Figure 4, the Masked Multi-Head Attention module will make the BERT model mask some characters during training. Let the Transformer model predict occluded characters, so that the model learns the features of the context and completes a cloze task. With the help of the contextual association features learned by the model, the BERT language model can learn the association between the learned labels in the substation label prediction task, and correctly predict the labels before and after. The prediction results of the ABINet-TS model are overly dependent on the prediction results of the first label in a series of labels, which may lead to the prediction results of subsequent texts being subject to the predictions of the first word. In order to solve the problem of contextual association of label text, we introduce the BERT model. implements a "cloze in the blank" style of contextual text line prediction reference.

4 Experiments

To meet the needs for identifying abnormal electrical cabinets during inspections, the experiments compare the accuracy of several state-of-the-art object detection models (MaskRCNN, PSENet, DBNet, and DBNet++) for detecting text boxes on a text detection dataset. For label recognition accuracy, the comparison includes cutting-edge models such as CRNN, SVTR, SATRN, MASTER, SAR, ASTER, ABINet, and ABINet-V on a single-label dataset. Finally, for text recognition accuracy on a concatenated image dataset, the models ABINet, ABINet-V, and ABINet-TS are compared.

4.1 Datasets and Implementation Details

Table. 1 Experimental Environment

Operating system	Ubuntu 18.04.1
GPU	NVIDIA GeForce RTX 2080Ti * 2
CPU	Intel(R)Core(TM)i9-9820X CPU
Technical Framework	Pytorch
Experiment Platform	JupyterLab
Coding Language	Python

Table. 2 Datasets

Dataset Type	Training images	Validation images
Text Detection Dataset	354	89
Single Label Dataset	3705	1040
Joined-Label Dataset	466	140

Table II shows that the text detection dataset is used to train the text detection models, enabling the models to locate the target text box positions from the images captured by the inspection robot. Each image in the single-label dataset contains a single text label. The concatenated image dataset is created by combining images from the single-label dataset, where the images are concatenated horizontally in an increasing order of label numbers. The images from the three datasets are shown in Figure 5.

The backbone network of the text detection model uses the Oclip pre-trained network with a training batch size of 8. The poly learning rate strategy is adopted, with an initial learning rate set to 0.007, power set to 0.9, weight decay set to 0.0001, and momentum set to 0.9.The data augmentation techniques applied during training include: (1)Random rotation within a range of ±10 degrees(2)Random cropping(3)Random flipping. All preprocessed images are resized to 640*640 to speed up training. During the inference phase, the aspect ratio of the test images is preserved.

The model channel size for the text recognition model is set to 512 for all layers.

In the BCN (Bidirectional Contextual Network), the four layers each have 8 attention heads.

The balancing factorsλ_v and λ_l are both set to 1, and the maximum prediction sequence length is set to 26. The images are resized to 32*128. The data augmentation techniques used during training include: (1)Geometric transformations (e.g., rotation, affine transformations) (2)Color transformations (3)Image quality degradation, among other methods. The batch size for the text recognition model is set to 384. The Adam optimizer is used, with an initial learning rate of e^{-4}, which decays to e^{-6}after 6 epochs.

In the training phase, the images from the concatenated image dataset are resized to 512 * 64. The model's maximum prediction sequence length is set to 80. The Xavier initialization is used for weight initialization, and the learning rate is adjusted accordingly during training.

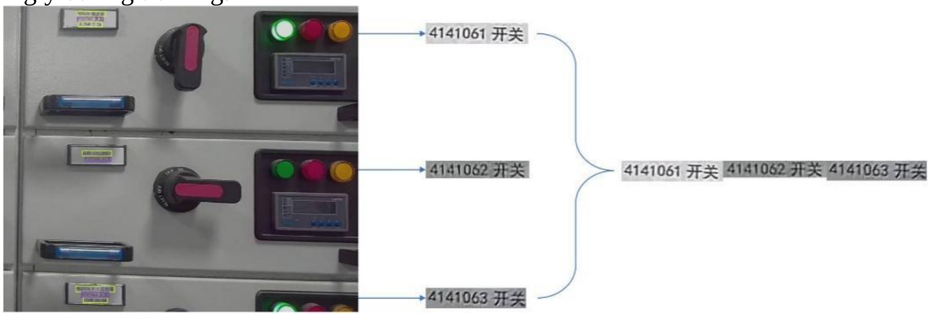

Fig. 5 Illustration of the three datasets. Left: Text detection dataset image, Middle: Single-label dataset image, Right: Concatenated label dataset image

5 Ablation Study

5.1 Text Detection

As shown in Table 3, the detection accuracy of DBNet++ for target bounding boxes reaches 0.9694, which is already sufficient for industrial scene requirements. DBNet uses a differentiable binarization method, which provides good detection results for the boundary regions of text labels. DBNet++ builds on DBNet by adding an adaptive scale fusion module, which effectively improves the model's robustness in detecting images at different scales. The detection accuracy of DBNet++ is now more than adequate to meet the needs of robotic inspections in substations.

Table. 3 Accuracy of the text detection models on the detection dataset

Models	Precision	Recall	Avg of precision and recall
MaskRCNN	0.8754	0.8931	0.8843
PSENet	0.8037	0.8462	0.8244
DBNet	0.9579	0.9635	0.9607
DBNet++	**0.9630**	**0.9760**	**0.9674**

Table. 4 Accuracy of the text recognition models on the single-label dataset

Models	Word Accuracy	Char Precision	Char Recall
CRNN	0.1923	0.7370	0.6584
SVTR	0.3865	0.8702	0.7998
SATRN	0.6587	0.8925	0.8845
MASTER	0.7385	0.9122	0.9053

SAR	0.7712	0.9152	0.9163
ASTER	0.8644	0.9663	0.9615
ABINet	<u>0.8933</u>	<u>0.9672</u>	<u>0.9619</u>
ABINet-V	**0.9058(+1.25%)**	**0.9750(+0.82%)**	**0.9683(+0.64%)**

Table 4 compares the recognition accuracy of different text recognition models on the single-label validation set. This table typically includes the accuracy metrics (such as Word Accuracy, Char Precision and Char Recall) for each model, which are used to evaluate how well each model performs in recognizing text labels on the validation set.

As shown in Table 4, ABINet-V achieves the highest recognition accuracy, with a character precision of 0.9750 and a text line accuracy of 0.9058. In contrast, the ABINet model, which incorporates a language model on top of ABINet-V, shows a decrease in recognition accuracy. This indicates that the semantic relationships between individual text labels are minimal, and the features extracted by the language model do not improve the predictions made by the visual model. Furthermore, since the loss function in ABINet includes losses from the visual model, the language model, and the fusion module, the visual model's predictions are influenced by the other loss functions, which may degrade its performance. This suggests that the visual model in ABINet performs well in recognizing the local information of text in images, while the language model excels at understanding contextual relationships between text labels. However, if the task goals for the visual and language models are not properly distinguished, it may lead to a "Crab Syndrome", where the performance is reduced due to conflicting objectives between the models.

In Figure 6, <UKN> represents 'UNKNOWN', indicating characters that the model failed to recognize. Among the models, ABINet-V delivers the best recognition performance, only mistaking PU275B as PU275, and demonstrates good performance in recognizing blurred labels.

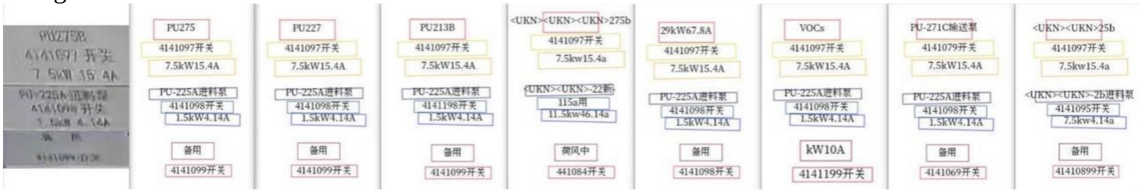

Fig. 6 The text recognition model experiment results comparison figure shows the following sequence from left to right: Original Image, ABINet-V, ABINet, ASTER, CRNN, MASTER, SAR, SATRN and SVTR recognition result

Table. 5 Oclip pretraining for ABINet-V and ABINet-TS

Models	Word Accuracy	Char Precision	Char Recall
ABINet-V	0.8957	0.9622	0.9676
ABINet-V(oclip)	0.9058	0.9750	0.9683
ABINet-TS	0.9122	0.9864	0.9875
ABINet-TS(oclip)	0.9174	0.9861	0.9892

Table. 6 Accuracy of the text recognition models on the joined-label dataset

Models	Word Accuracy	Char Precision	Char Recall
ABINet-V	0.8440	0.9801	0.9833
ABINet	0.8624	0.9834	0.9866
ABINet-TS	**0.9174(+4.5%)**	**0.9844(+0.1%)**	**0.9877(+0.11%)**
ABINet-TSB	**0.9216(+4.92%)**	**0.9861(+0.33%)**	**0.9892(+0.26%)**

Table VI compares the recognition accuracy of ABINet-V, ABINet, and ABINet-TS on the joined-label image validation set. From Table VI, it can be observed that ABINet-V, which performs better on single-label images, actually shows lower accuracy on concatenated images compared to ABINet. This indicates that the language model in ABINet is effective in recognizing the relationships between text labels. The ABINet-TS model, by separating the tasks of the visual model and the language model, allows both to focus more effectively on their respective recognition tasks. Specifically, the visual model extracts features and predicts text for individual images, while the language model corrects the predictions based on the context. The visual model focuses on recognizing local text information, and the language model focuses on the relationship between individual recognition results. These are two distinct task goals for each layer. Compared to ABINet-V in Table IV, ABINet-TS shows an improvement of 1.44\% in character precision and character recall, and a 1.16% increase in text line accuracy. This suggests that the separation of tasks in ABINet-TS leads to better overall performance when handling concatenated images. Compared with the ABINet-TSB model, the accuracy of the ABINet-TSB model is improved by 0.42%, which indicates that the BERT model improves the ability to extract contextual features, so that the ABINet-TS model is not affected by the pre-predicted text characters.

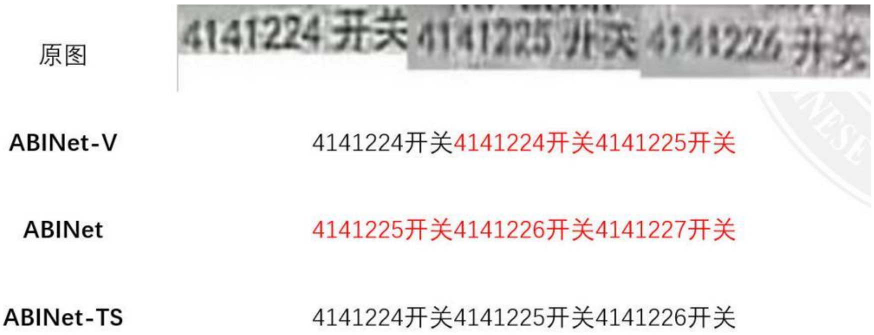
原图
4141224开关4141225开关4141226开关
ABINet-V
4141224开关4141224开关4141225开关
ABINet
4141225开关4141226开关4141227开关
ABINet-TS
4141224开关4141225开关4141226开关

Fig. 7 The recognition results of three models

As depicted in Figure 7, All individual labels ABINet predict are incorrect. This indicates that ABINet's language model, while considering the relationships between the labels, does not perform well on concatenated images. The autoregressive nature of the language model can cause errors from earlier predictions to propagate and affect the recognition of subsequent characters. ABINet-V: Only the first label is correctly recognized. This demonstrates that ABINet-V focuses more on local character prediction and does not take into account the semantic relationships between the labels. It works well for individual labels but struggles when those labels are concatenated into a single image. ABINet-TS: ABINet-TS ensures that the visual model's local predictions are not disturbed by the language model. It successfully handles the recognition task by allowing the visual model to focus on extracting features for individual labels, while the language model uses context to correct the predictions. The final result shows the correct label combinations, as the language model is able to integrate the semantic information from the visual model's predictions. This comparison highlights the importance of task separation in ABINet-TS, which balances the strengths of both the visual and language models, resulting in more accurate recognition of concatenated images. ABINet and ABINet-V show limitations in handling multi-label images, as their models either lack contextual information (ABINet-V) or suffer from error propagation due to the autoregressive mechanism (ABINet).

6 Conclusion

Accurate recognition of electrical cabinet labels is essential for intelligent recording of inspection anomalies. In this paper, we applied DBNet++ for precise detection of abnormal electrical cabinet label locations, achieving a text line recognition accuracy of 89.33\% using ABINet. With the help of pre-training of the oclip model, the visual model of ABINet improves the text line recognition accuracy to 90.58\%, indicating that the visual model is proficient at recognizing local text information in images. The language model in ABINet, on the other hand, excels at recognizing the contextual relationships between the text. Both models are constrained by a joint loss function, which balances their contributions. To address identified issues, we propose a distributed hierarchical structure for the multi-objective OCR text recognition model framework, ABINet-TS. With this framework, the text line recognition accuracy reaches 91.74\%, better meeting the needs of robotic intelligent inspection tasks for the automatic reading and recording of electrical cabinet labels. The language model in ABINet-TS was replaced with BERT, which further improved the accuracy of text lines to 92.16\%. This improvement ensures more reliable performance in the robotic inspection process, enabling efficient automation in labeling and anomaly detection within electrical substations.

Acknowledgments. This work was supported by the National Science Foundation of China under Grant 62333015, and Grant 61673366. (Corresponding author: Shu Lin.)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

2. Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010.
3. Neumann, Lukáš, and Jiří Matas. "Real-time scene text localization and recognition." 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012.
4. Busta, M., L. Neumann, and J. Matas. "Fastext: Effffcient unconstrained scene text detector, 2015." IEEE International Conference on Computer Vision (ICCV). Vol. 1. No. 9.
5. Liao, Minghui, et al. "Textboxes: A fast text detector with a single deep neural network." Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017.
6. Liao, Minghui, Baoguang Shi, and Xiang Bai. "Textboxes++: A single-shot oriented scene text detector." IEEE transactions on image processing 27.8 (2018): 3676-3690.
7. He, Pan, et al. "Single shot text detector with regional attention." Proceedings of the IEEE international conference on computer vision. 2017.
8. Zhou, Xinyu, et al. "East: an efficient and accurate scene text detector." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017.
9. Zhang, Zheng, et al. "Multi-oriented text detection with fully convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
10. Xue, Chuhui, Shijian Lu, and Fangneng Zhan. "Accurate scene text detection through border semantics awareness and bootstrapping." Proceedings of the European conference on computer vision (ECCV). 2018.
11. Wang, Wenhai, et al. "Shape robust text detection with progressive scale expansion network." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
12. Liao, Minghui, et al. "Real-time scene text detection with differentiable binarization and adaptive scale fusion." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 919-931.
13. Tang, Jun, et al. "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping." Pattern recognition 96 (2019): 106954.
14. Shi, Baoguang, Xiang Bai, and Cong Yao. "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." IEEE transactions on pattern analysis and machine intelligence 39.11 (2016): 2298-2304.
15. Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Proceedings of the 23rd international conference on Machine learning. 2006.
16. He, Pan, et al. "Reading scene text in deep convolutional sequences." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.
17. Hu, Wenyang, et al. "Gtc: Guided training of ctc towards efficient and accurate scene text recognition." Proceedings of the AAAI conference on artiffcial intelligence. Vol. 34. No. 07. 2020.
18. Du, Yongkun, et al. "Svtr: Scene text recognition with a single visual model." arXiv preprint arXiv:2205.00159 (2022).
19. Yu, Deli, et al. "Towards accurate scene text recognition with semantic reasoning networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
20. Lee, Chen-Yu, and Simon Osindero. "Recursive recurrent nets with attention modeling for ocr in the wild." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
21. Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

22. Cheng, Zhanzhan, et al. ”Focusing attention: Towards accurate text recognition in natural images.” Proceedings of the IEEE international conference on computer vision. 2017.
23. Cheng, Zhanzhan, et al. "Aon: Towards arbitrarily-oriented text recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
24. Wojna, Zbigniew, et al. "Attention-based extraction of structured information from street view imagery." 2017 14th IAPR international conference on document analysis and recognition (ICDAR). Vol. 1. Ieee, 2017.
25. Lyu, Pengyuan, et al. "2d attentional irregular scene text recognizer." arXiv preprint arXiv:1906.05708 (2019).
26. Wang, Yuxin, et al. "From two to one: A new scene text recognizer with visual language modeling network." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
27. Atienza, Rowel. "Vision transformer for fast and efficient scene text recognition." International conference on document analysis and recognition. Cham: Springer International Publishing, 2021.
28. Du, Yongkun, et al. "Svtr: Scene text recognition with a single visual model." arXiv preprint arXiv:2205.00159 (2022).
29. Fang, Shancheng, et al. "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
30. Shi, Baoguang, et al. "Aster: An attentional scene text recognizer with flexible rectification." IEEE transactions on pattern analysis and machine intelligence 41.9 (2018): 2035-2048.
31. Xue, Chuhui, et al. "Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.