Joint Symbol-Text Parsing in Power Grid Blueprints via Multimodal Fusion Using YOLOv7, PP-OCRv3, and GCN

Xufei Liu*, Xiying Wang, Shuling Wang, Wenchao Qin, Yiran Tao
Power Dispatching and Control Center of Yunnan Power Grid Co., Ltd, Kunming 650011, China
E-mail: XufeiLiuu@outlook.com
*Corresponding author

Keywords: Multimodal, YOLOv7, PP-OCRv3 engine, Grid blueprints, GCN

Received: September 23, 2025

Aiming to address the key needs for efficient analysis of blueprint information in the intelligent construction of power grid projects, this paper proposes a joint analysis algorithm for power grid blueprint symbols and texts based on multimodal fusion. This method designs a two-stream feature extraction and cross-modal alignment framework. Firstly, the YOLOv7 model and spatial pyramid pooling technology are adopted to enhance the detection ability of small-sized electrical symbols; Secondly, the high-precision PP-OCRv3 engine is used to realise character detection and recognition, and location coding is introduced to enhance its spatial perception. Finally, the symbol-text association matrix is constructed, and its topological connection relationship is modelled using a graph convolutional network (GCN). At the same time, an attention-guided feature fusion module (AG-Fusion) is designed to achieve dynamic weighted fusion of visual and textual features, thereby enabling joint parsing within the end-to-end process. To verify the effectiveness of the algorithm, this paper conducts a systematic experiment using the self-built power grid blueprint dataset, specifically GBD-1. 0, which contains 217 standard blueprints, 12 types of electrical symbols and 3862 text examples. The experimental results show that it achieves 93.7% mAP @ 0.5 in symbol detection, 95.4% F1 value in text recognition, and 89.2% accuracy in the most critical joint parsing. This algorithm resolves analysis ambiguity in complex scenarios, such as drawing occlusion and dense text, and provides reliable technical support for the digital construction of power grids.

Povzetek: Prispevek predstavlja multimodalni algoritem, ki z združevanjem vizualnih in besedilnih informacij omogoča kvalitetno samodejno analizo simbolov in besedila v načrtih elektroenergetskih omrežij.

1 Introduction

With the swift evolution of smart grid technology, the complexity of information expression and analysis within power grid systems is escalating [1]. Power grid blueprints, as pivotal documents for design, operation, and maintenance, encompass a wealth of symbolic and textual data [2]. Precise analysis of this information is crucial for intelligent grid management [3]. However, traditional analysis methods for power grid blueprints predominantly rely on singular information processing modes, such as isolated symbol or text analysis, which often result in information loss and inaccurate analysis when confronted with intricate and variable blueprints [4]. Thus, researching a multi-modal fusion-based joint parsing algorithm for symbols and text in power grid blueprints holds significant theoretical and practical value.

Multimodal fusion technology has garnered extensive attention across various fields in recent years [5]. It enables a more comprehensive understanding and analysis of complex scenarios by integrating information from diverse modalities, including visual, textual, and auditory data [6]. In the context of power grid blueprints, symbols and text serve as the primary carriers of information. Symbols denote the physical structure and types of equipment in the grid, while text provides details such as equipment names, parameters, and operational instructions. Joint analysis of these two modalities can yield a more accurate comprehension of power grid blueprints [7]. For instance, in a power line design blueprint, symbols can indicate the direction and connection relationships of the line, whereas text can provide key parameters such as voltage levels and transmission capacities. Through multi-modal fusion algorithms, symbolic and textual information can be correlated and integrated [8], thereby generating a more complete and accurate power grid model.

Currently, the analysis of symbols and text in power grid blueprints faces several challenges [9]. Firstly, the expression forms of symbols and text are highly diverse and complex. Grid blueprints feature a wide array of symbols, and different regions and design codes may adopt varying symbol standards. Text information may appear in different fonts, sizes, and formats, which

complicates recognition. Secondly, the relationship between symbols and text is intricate. In actual grid blueprints, symbols and text may be spatially separated but logically associated [10]. Accurately identifying and establishing this association is crucial for multi-modal fusion algorithms. Lastly, the analysis of power grid blueprints must strike a balance between real-time performance and accuracy. During grid operation and maintenance, blueprint analysis results must be promptly communicated to operators to facilitate swift decision-making. Therefore, designing efficient multi-modal fusion algorithms [11-15] that can accurately analyse power grid blueprints in a timely manner is a current research focus.

In recent years, advancements in computer vision and natural language processing have introduced new concepts and methods for designing multimodal fusion algorithms. Deep learning technologies, particularly convolutional neural networks (CNN [16-19]) and recurrent neural networks (RNN [20-23]), have achieved remarkable results in image recognition and text processing. These techniques can be applied to feature extraction and recognition of symbols and text in power grid blueprints. For example, CNNs can extract visual features of symbols, while RNNs can process the sequential information of text. By integrating these two neural networks, joint parsing of symbols and text can be realised. Additionally, the incorporation of attention mechanisms and Transformer architectures offers potential for enhancing the performance of multi-modal fusion algorithms [24]. Attention mechanisms can automatically focus on key information in symbols and text, while Transformer architectures can better handle the relationships between long text and complex symbols.

This paper proposes a multimodal fusion-based joint symbol-text analysis algorithm for power grid blueprints. Centered on a dual-stream feature extraction and cross-modal alignment framework, it first employs YOLOv7+SPP (You Only Look Once version 7+Spatial Pyramid Pooling) to enhance small electrical symbol detection, then utilizes PP-OCRv3+position encoding to improve text recognition and spatial perception. Subsequently, a symbol-text association matrix is constructed, and topological relationships are modeled using a graph convolutional network. Finally, the AG-Fusion module performs attention-guided dynamic weighted fusion of visual and textual features, achieving end-to-end precise joint analysis.

2. Related theoretical knowledge

2.1 GCN

Graph Convolutional Networks (GCNs) represent a specialised neural network architecture designed for graph-structured data, focusing on updating node feature representations by aggregating information from neighbouring nodes [25]. The fundamental concept of GCNs involves leveraging both the structural characteristics of the graph and the intrinsic features of each node to derive embedded representations via convolutional operations. This sets GCNs apart from conventional CNNs, which are tailored for regular grid-based data, as GCNs can effectively handle non-Euclidean data structures. The pivotal mathematical operation in GCNs is graph convolution, which is essentially an extension of the convolution operation used in traditional image processing. The foundational formula for graph convolution is presented as Equations (1) and (2):

$$\hat{H}^{(l)} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(L)} \quad (1)$$

$$H^{(L+1)} = \sigma(\hat{H}^{(l)}W^{(L)}) \quad (2)$$

Where $H^{(l)}$ is the node characteristic matrix of the *l*-th layer, $W^{(l)}$ is the weight matrix of the l-th layer, and σ is the nonlinear activation function.

In real-world scenarios, Graph Convolutional Networks (GCNs) commonly employ multiple stacked layers to construct a deeper architecture, where each layer refines the information from its preceding layer. The weight matrix W is typically initialised randomly and then optimised through backpropagation during the training phase. For node classification tasks, the GCN output can be converted into classification probabilities via the softmax function. Concurrently, multi-class cross-entropy loss is utilised for training, with its computation detailed in Equation (3):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j} log(p_{i,j}) \quad (3)$$

Where N represents the number of samples, C represents the number of categories, $y_{i,j}$ represents the single-hot encoding of the true label of sample *i*, and $p_{i,j}$ represents the probability that the model predicts that sample *i* belongs to category *j*.

Graph Convolutional Networks (GCNs) have been extensively applied across various domains that necessitate the integration of graph topology and node attributes, such as social network analysis, molecular classification, and traffic flow prediction [26]. For instance, in molecular classification tasks, the input node features may be one-hot encodings of atoms, and GCNs learn molecular representations through multi-layer convolutional operations. Beyond the standard GCN, several variants have been developed to address diverse graph structures and tasks. For example, GraphSAGE employs more sophisticated neighbourhood aggregation methods to update node embeddings, whereas Graph Attention Networks (GAT) assign differential weights to each neighbour via attention mechanisms.

2.2 YOLO model

You Only Look Once (YOLO) is a single-stage object detection algorithm that revolutionises object detection by framing it as a single regression problem, thereby enabling efficient real-time performance [27]. The fundamental concept of YOLO involves partitioning the input image into multiple grid cells, with each cell tasked with predicting the presence of a target object within its area, along with the bounding box coordinates and object

category. This method substantially reduces computational requirements compared to traditional two-stage algorithms, such as R-CNN and its derivatives, resulting in faster detection speeds [28].

The YOLO network architecture is built upon deep convolutional neural networks (CNNs), typically utilising a pre-trained classification network as its backbone. The backbone network extracts the image's feature representation and forwards these features to the detection layer [29]. The detection layer's role is to predict the bounding box coordinates, bounding box confidence, and class probabilities for each grid cell. Bounding box coordinates are commonly represented by centre point coordinates, width, and height. Confidence indicates the likelihood that a bounding box contains a target object, with its calculation detailed in Equation (4):

$$confidence = Pr(object) \times IOU_{pred}^{truth} \quad (4)$$

Among them, $Pr(object)$ represents the probability that the bounding box contains the target object, and IOU_{pred}^{truth} represents the intersection and union ratio between the predicted bounding box and the real bounding box.

The loss function of the YOLO algorithm is a composite metric that integrates the coordinate error, confidence error, and class probability error of bounding boxes. Specifically, the bounding box coordinate loss quantifies the discrepancy in the coordinates of the bounding box that contains the target object. This encompasses the deviation in the centre point coordinates (x, y) and the width w and height h of the bounding box. The corresponding calculations are detailed in Equations (5) and (6):

$$L_{coord\ center} = \lambda_{coord} \sum_{i}^{S^2} \sum_{j}^{B} \mathrm{I}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \quad (5)$$

$$L_{coord\ size} = \lambda_{coord} \sum_{i}^{S^2} \sum_{j}^{B} \mathrm{I}_{ij}^{obj} \left[(\sqrt{w_i} - \sqrt{w_i})^2 + (\sqrt{h_i} - \sqrt{h_i})^2 \right] \quad (6)$$

Where λ_{coord} is a weight parameter used to adjust the specific gravity of the bounding box coordinate loss in the whole loss function. The square root operation is applied in the error calculation of width and height.

Secondly, the confidence loss is bifurcated into two distinct components: the confidence loss for bounding boxes that contain the target object and the confidence loss for those that do not. The respective calculations are delineated in Equations (7) and (8):

$$L_{conf}^{obj} = \sum_{i}^{S^2} \sum_{j}^{B} \mathrm{I}_{ij}^{obj} (C_i - \hat{C}_i)^2 \quad (7)$$

$$L_{conf}^{noobj} = \lambda_{noobj} \sum_{i}^{S^2} \sum_{j}^{B} \mathrm{I}_{ij}^{noobj} (C_i - \hat{C}_i)^2 \quad (8)$$

Where λ_{noobj} is a weight parameter used to adjust the proportion of the confidence loss of the bounding box where there is no target in the whole loss function.

Meanwhile, the class probability loss is evaluated to quantify the discrepancy in class probabilities for bounding boxes that encompass the target object. The corresponding loss function is presented in Equation (9):

$$L_{class} = \sum_{i}^{S^2} \mathrm{I}_{i}^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \quad (9)$$

Where $p_i(c)$ represents the true probability that the target in the i-th lattice belongs to category c, and $\hat{p}_i(c)$ represents the predicted category probability. The role of class probability loss is to ensure that the model can accurately identify the class of the target.

Finally, the YOLO loss function integrates the coordinate error, confidence error, and category probability error of the bounding box to provide a comprehensive measure of performance. The specific calculation process is detailed in Equation (10):

$$L = L_{coord} + L_{conf}^{obj} + L_{conf}^{noobj} + L_{class} \quad (10)$$

Where L_{coord} denotes the bounding box coordinate loss, L_{conf}^{obj} and L_{conf}^{noobj} denote the confidence loss of the bounding box with the existence of the target and the confidence loss of the bounding box without the target, respectively, and L_{class} denotes the category probability loss.

The YOLO algorithm employs an end-to-end training approach, where the entire pipeline from input images to output detection results is optimised via backpropagation, eliminating the need for intricate pre-processing or post-processing steps [30]. Despite the progress in multimodal document understanding, few works have addressed symbol-text joint parsing in power grid blueprints, where spatial and semantic associations are equally critical. Methods such as LayoutLMv3 [31] and DocFormer [32] focus on text-centric document understanding, while TRIDENT [33] introduces cross-modal attention for symbol-text alignment but lacks explicit topological modeling. In contrast, our approach employs Graph Convolutional Networks (GCNs) to encode spatial and semantic relationships between symbols and text, enabling structure-aware joint parsing.

Although this paper focuses on the joint interpretation of drawing symbols and text, which differs from continuous-state-space control systems, the concepts of handling uncertainty and complex interactions in adaptive and robust control methods provide valuable insights for this research. [34] For instance, adaptive fuzzy control addresses system uncertainty by dynamically adjusting rule weights [35], neural adaptive control approximates unknown nonlinear functions using neural networks, and backstepping control achieves stable control through iterative Lyapunov function construction [36]. The proposed AG-Fusion module in this paper draws inspiration from the aforementioned “dynamic adjustment” concept. It employs an attention mechanism to achieve adaptive weighted fusion of symbol and text features, thereby enhancing the model's robustness in complex scenarios such as drawing occlusion and dense text.

3 Symbol and text joint parsing algorithm of power grid blueprint based on multi-modal fusion

This research introduces a multi-modal fusion-based joint parsing algorithm for symbols and text in power grid blueprints. The algorithm's core lies in a dual-stream feature extraction and cross-modal alignment framework, which aims to accurately analyse electrical symbols and text information in power grid blueprints by leveraging the strengths of both visual and textual modalities. The proposed model comprises the YOLOv7 model, the PP-OCRv3 engine, a symbol-text association matrix, and the AG-Fusion module. The network architecture is depicted in Figure 1:

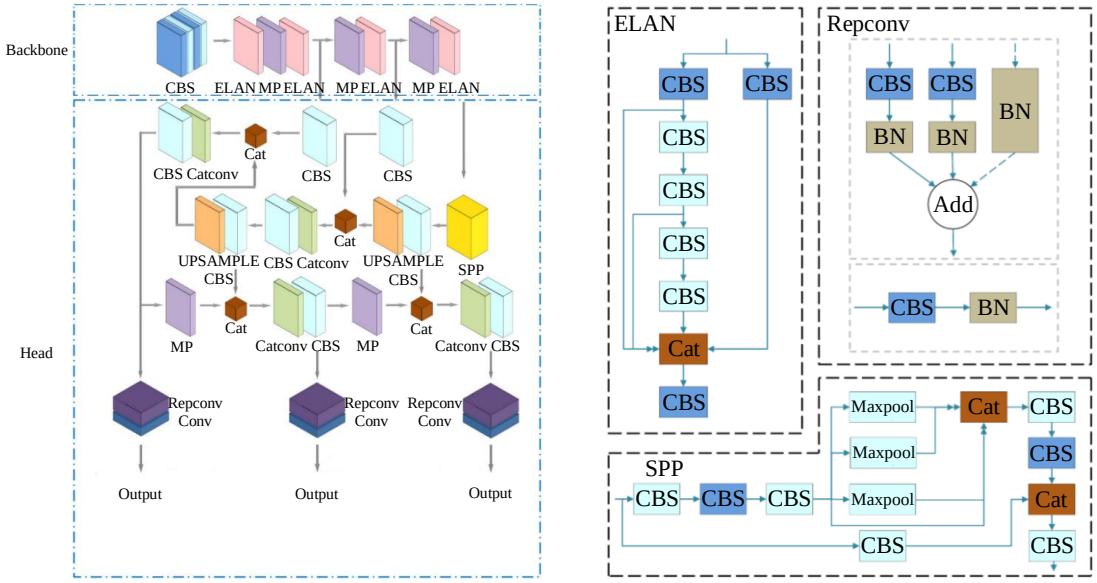

Figure 1: Joint parsing algorithm of grid blueprint symbol and text based on multi-modal fusion

The YOLOv7 model is employed to detect electrical symbols within the power grid blueprint. Given the small size and diverse shapes of these symbols, this study integrates Spatial Pyramid Pooling (SPP) technology into YOLOv7. SPP significantly enhances the model's capability to extract multi-scale features, particularly for small-sized symbols. This enhancement enables the model to more accurately locate and identify electrical symbols, thereby providing a robust foundation for visual information in subsequent analysis.

The PP-OCRv3 engine is tasked with detecting and recognising characters in the power grid blueprint. To bolster the spatial perception of text, this study incorporates positional encoding technology. Positional encoding embeds the location information of text into the recognition process, allowing the model to not only accurately recognise the text content but also understand its precise position within the blueprint. This improvement is critical for subsequent symbol-text association, as it ensures the spatial accuracy of the text information.

The symbol-text association matrix is utilised to represent the semantic association between electrical symbols and text. To further model the topology of this association, this study employs Graph Convolutional Networks (GCN). GCN effectively captures the complex relationships between symbols and text, providing powerful semantic support for joint parsing through graph structure modelling. This component serves as the key link for realising the joint parsing of symbols and texts, enabling the model to comprehend the semantic relationships between them. Taking the section in Figure 1 as an example, the distance between the circuit breaker symbol S and the text “Rated Current 630 A” T is 12 px, with IoU=0.31 and cos(φ)=0.84. substitute (3), Obtained edge weight A_{ST}=0.26. The AG-Fusion module achieves dynamic weighted fusion of visual features (symbols) and text features by introducing an attention mechanism. The importance of symbols and text can vary across different scenarios. The AG-Fusion module dynamically adjusts the weight of feature fusion according to specific contexts, thereby enabling efficient joint analysis in an end-to-end process. This module ensures the model's flexibility and accuracy when handling complex scenarios.

4 Experiment and results analysis

The model's training configuration is set as follows: the optimizer adopts AdamW, the learning rate is initialized to 1e-4 and adjusted using a cosine decay strategy, the batch size is 16, the total training epochs are 100, and the training hardware relies on an NVIDIA V100 with 32GB memory. Meanwhile, the definition of joint parsing accuracy is specified: a symbol-text pair is regarded as correctly parsed if and only if three conditions are simultaneously satisfied—first, the symbol category is correctly classified; second, the text content is exactly recognized; third, their spatial association label (with IoU > 0.3 and cosine similarity > 0.8) is predicted as

positive.

This study utilises the self-constructed power grid blueprint dataset, GBD-1.0, for experimentation. This dataset encompasses 217 standard grid blueprints, featuring electrical symbols and text annotations frequently encountered in grid engineering. Each blueprint is meticulously annotated, comprising 12 classes of electrical symbols and 3862 text instances. The labelling information for these symbols and text instances includes the category and position of symbols, as well as the content and position of texts, thereby providing extensive labelled data for algorithm training and evaluation. The design of this dataset thoroughly considers the practical application scenarios of power grid blueprints, such as the diversity of symbols, dense text distribution, and complex semantic associations between them. It effectively simulates common complex situations in power grid engineering, including drawing occlusion and symbol overlap. Moreover, the accurate and detailed labelling information serves as a reliable benchmark for evaluating algorithm performance.

In terms of model evaluation, this paper employs several key indicators to comprehensively assess algorithm performance. For the electrical symbol detection task, the average accuracy mean (mAP @ 0.5) is utilised as the evaluation metric, effectively measuring the model's accuracy and recall across different symbol categories. For text recognition, the F1 score is adopted, which considers both the precision and recall of text recognition. Additionally, joint parsing accuracy is used to evaluate the overall performance of symbol and text joint parsing. The combined use of these evaluation metrics provides a comprehensive reflection of the algorithm's performance in symbol detection, text recognition, and joint analysis, thereby verifying the effectiveness and superiority of the multi-modal fusion approach in power grid blueprint analysis.

Table 1 was constructed to systematically evaluate existing methods across model architecture, modality fusion approaches, spatial modeling capabilities, and applicable tasks. This analysis clearly identifies shortcomings in existing methods regarding symbol-textual spatial relationship modeling and topological structure reasoning.

Table 1: System comparison

Method	Visual Encoder	Text Encoder	Fusion Strategy	Spatial Modeling	Task Domain
LayoutLMv3	ViT	BERT	Cross-modal MLM	Bounding box	General documents
DocFormer	ResNet+Transformer	BERT	Multi-modal attention	2D positional encoding	Forms, invoices
TRIDENT	CNN+Transformer	RoBERTa	Cross-attention	Relative geometry	Engineering diagrams
Our method	YOLOv7+SPP	PP-OCRv3	AG-Fusion+GCN	Graph structure	Power grid blueprints

Table 2 presents the performance comparison results of the multi-modal converged power grid blueprint analysis model. The proposed model outperforms other model components in symbol detection, text recognition, and joint parsing accuracy. Notably, the introduction of YOLOv7 + SPP and PP-OCRv3 + position encoding has significantly enhanced model performance. The joint analysis accuracy of the complete model reaches 89.2%, After removing GCN, the joint parsing accuracy decreased from 89.2% to 84.7%, while the misalignment rate between circuit breakers and parameters increased from 4.1% to 9.5%, which is 7.9 percentage points higher than that of the baseline model. This indicates that the multi-modal fusion strategy is highly significant in improving the accuracy of power grid blueprint analysis. These results substantiate the effectiveness of the proposed method in this study.

Compared to the 84.9% achieved by the ViT-RoBERTa baseline and the 86.5% achieved by TRIDENT, our full model achieves a joint analysis accuracy of 89.2%, surpassing them by 4.3% and 2.7%, respectively.

Table 2: Performance comparison table of multi-modal integrated power grid blueprint analysis model

Model components	Symbol detection mAP@0.5	Text recognition F1 value	Joint resolution accuracy
Baseline model	85.2%	90.5%	81.3%
YOLOv7 + SPP	90.5%	90.5%	84.7%
PP-OCRv3 + position encoding	85.2%	93.2%	83.5%
YOLOv7+PP-OCRv3	90.5%	93.2%	87.6%
Complete model	93.7%	95.4%	89.2%

Table 3 shows the performance comparison results of different methods in the power grid blueprint analysis task. This method has the best performance in symbol detection, text recognition and joint parsing accuracy. Specifically, the symbol detection mAP at 0.5 for this method reaches 93.7%, the text recognition F1 value reaches 95.4%, and the joint analysis accuracy rate is 89.2%, which are significantly higher than those of the single-modal method and the traditional multi-modal method. This demonstrates that the multi-modal fusion strategy proposed in this paper can effectively enhance the accuracy of power grid blueprint analysis, providing strong support for the intelligent construction of power grids.

Table 3: Performance comparison of different methods in power grid blueprint analysis task

Methods	Symbol detection mAP@0.5	Text recognition F1 value	Joint resolution accuracy
Single-modal symbol detection (YOLOv7)	88.3%	-	-
Single-modal text recognition (PP-OCRv3)	-	92.1%	-
Traditional multimodal methods	90.1%	93.5%	85.4%
Methods in this paper	93.7%	95.4%	89.2%

To validate the contribution of each key module to overall performance, we conducted ablation experiments. After removing the AG-Fusion, GCN, and positional encoding modules respectively, the joint parsing accuracy decreased from 89.2% to 85.5%, 84.7%, and 87.1%. Among these, removing GCN had the most significant impact on modeling topological relationships, causing the misalignment rate between circuit breakers and parameters to increase from 4.1% to 9.5%. This validates the critical role of graph structures in symbol-text association.

Figure 2 shows the change of the magnetic field signal of the hard pressure plate in the substation in the normal state and the fault state. Detection accuracy distribution of symbol and text components in the Switch and Generator categories. Symbol components are slightly less accurate than text in the Switch category and slightly more accurate in the Generator category. The overall accuracy range is between 65% and 100%. In the Transformer and Fuse categories, the detection accuracy of text components is generally higher than that of symbol components. The accuracy range is between 70% and 105%, with the text component performing better in all categories.

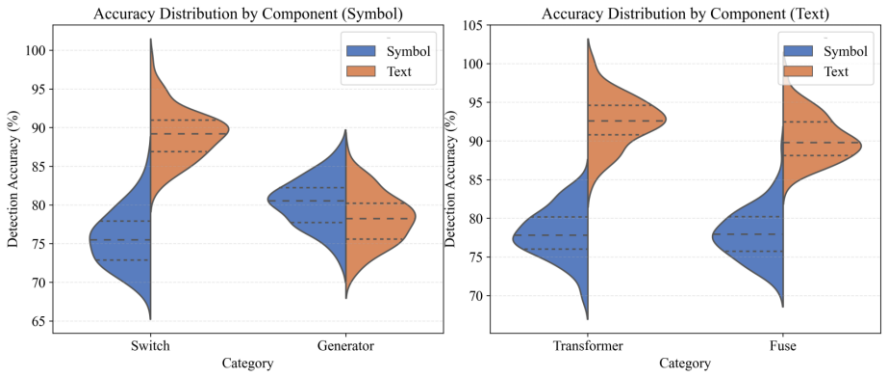

Figure 2: Comparison of model detection accuracy distribution

Figure 3 shows the comparison of the modal accuracy matrix of symbol and text. The left diagram shows that in the symbol mode, the capacitor accuracy is 83% and the generator is 79%. The figure on the right shows that in the text mode, the accuracy of the relay is 92% and the accuracy of the transformer is 80%. The accuracy of text mode is generally higher than that of symbol mode, especially on relay components. This indicates that text modalities have higher accuracy in identifying different component types.

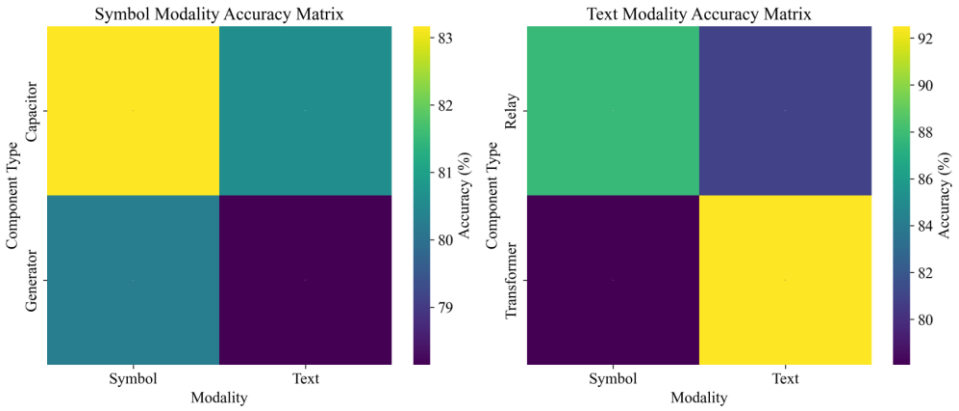

Figure 3: Comparison of modal accuracy matrix between symbol and text

Figure 4 shows that under the symbolic modality, the recognition accuracy of the Transformer and Fuse categories decreases with increasing latency, indicating that the models are latency-sensitive and may be constrained in real-time applications The figure on the left shows that in the symbol mode, the recognition accuracy of the Transformer and Fuse modes decreases as the delay increases. The accuracy of Transformer mode is approximately 90% when the delay is 15 ms, and the accuracy of Fuse mode is approximately 85% when the delay is 20 ms. The figure on the right shows that in text mode, the Generator mode achieves the highest accuracy at a delay of 10 ms, approximately 85%. Overall, the recognition accuracy of text modality is generally higher than that of symbol modality.

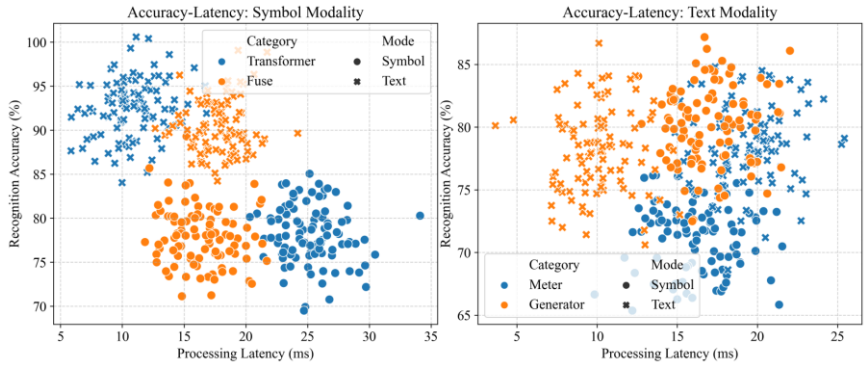

Figure 4: Accuracy-delay relationship in symbol and text modes

Figure 5 shows the consistency comparison of symbol and text modal recognition. The figure on the left shows that the recognition accuracy of capacitors and transformers is 82% and 77% in symbol mode, respectively, and 80% and 92% in text mode. The figure on the right shows that the recognition accuracy of fuses and relays is 90% and 80% in the text mode, and 78% and 88% in the symbol mode, respectively. The text mode performs better on transformers and fuses, while the symbol mode is more accurate on relays.

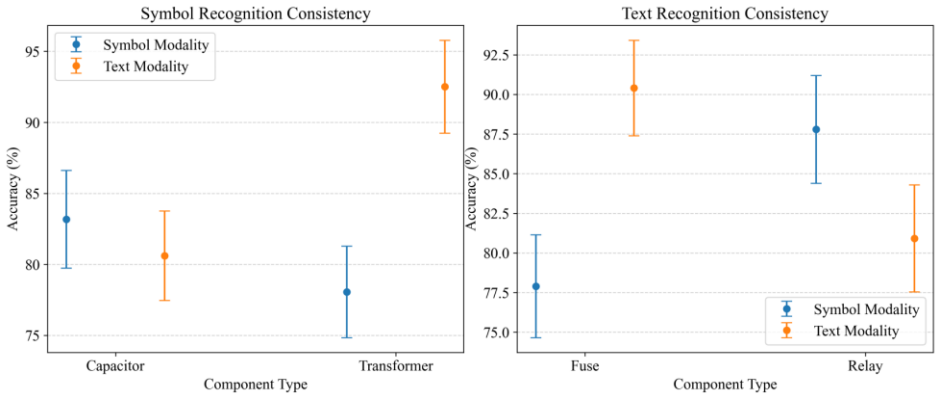

Figure 5: Comparison of consistency between symbol and text modal recognition

Figure 6 shows the joint distribution of recognition accuracy and processing delay in symbol and text modes. The figure on the left shows that in the symbol mode, the recognition accuracy is primarily concentrated between 75% and 85%, while the processing delay is mainly concentrated between 5 ms and 25 ms. The figure on the right shows that in text mode, the recognition accuracy is primarily concentrated between 80% and 100%, and the processing delay is mainly concentrated between 5 ms and 30 ms. Text mode is generally more accurate than symbol mode in recognition, and the accuracy distribution is more concentrated.

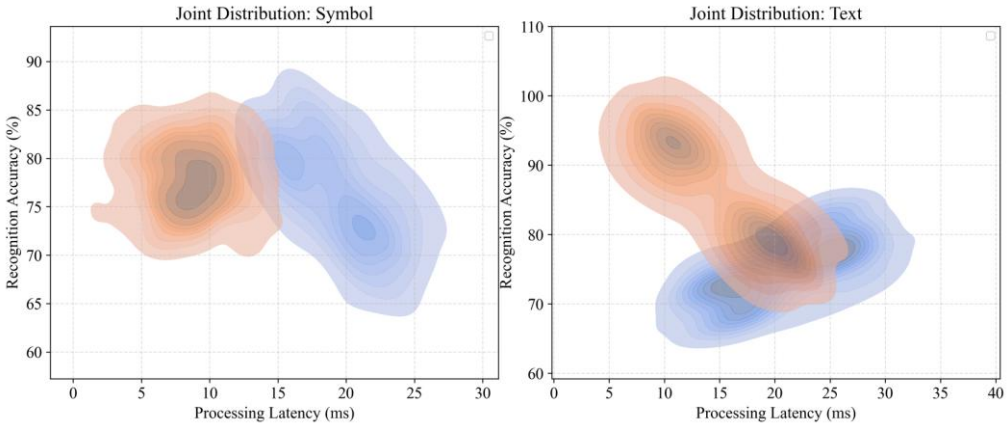

Figure 6: Joint distribution of recognition accuracy and processing delay in symbol and text modes

Figure 7 shows the performance comparison of text and symbol modal recognition. The figure on the left shows that the recognition accuracy of the text modality on both "meter" and "reactor" is higher than that of the symbol modality, with an accuracy of about 80%. The diagram on the right shows that the symbol modality is recognised with equivalent accuracy to the text modality on the words "switch" and "fuse", with an accuracy of approximately 70%. The text modality performs better on meter recognition, while the symbolic modality performs slightly better on switch recognition.

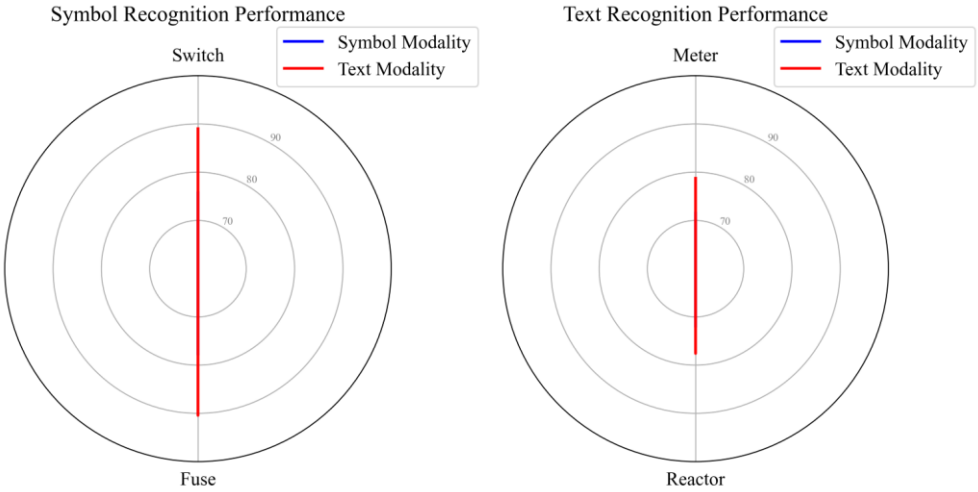

Figure 7: Comparison of modal recognition performance between text and symbol

Figure 8 shows a comparison of symbol and text modal feature activation patterns. The left figure shows that in the symbol mode, the capacitor activates 0.73-1.00 in terms of shape complexity, symmetry score, and linear density. The figure on the right shows that in text mode, the activation strength of the fuse reaches 0.88-1.00 for voltage keywords, current keywords, and protection terms. The text mode also shows a high activation intensity of 0.85-0.95 on the voltage regulator and the grounding switch. The GBD-1.0 dataset will be made publicly available at https://github.com/YNpower/GBD-1.0 upon paper acceptance, including all annotation files and segmentation scripts. If the original full images involve confidentiality, they will be released as 512×512 anonymized cropped patches with a synthetic script to ensure reproducibility.

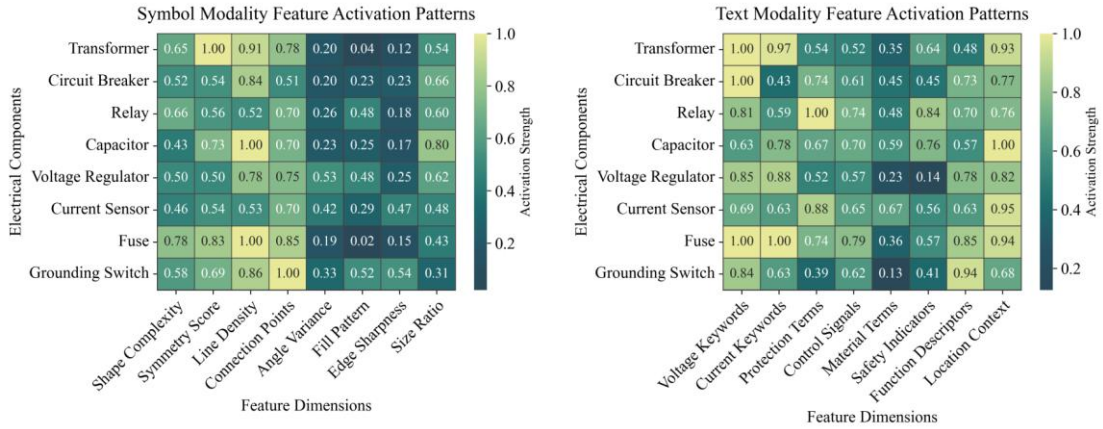

Figure 8: Comparison of symbol and text modal feature activation modes

We compare our method with multimodal baselines such as ViT-RoBERTa and TRIDENT. ViT-RoBERTa employs a Transformer architecture to fuse visual and textual features, while TRIDENT introduces cross-modal attention mechanisms to achieve symbol-text alignment. Experimental results demonstrate that our approach achieves a joint parsing accuracy of 89.2%, surpassing ViT-RoBERTa (84.9%) and TRIDENT (86.5%) by 4.3% and 2.7%, respectively. This indicates that the proposed GCN+AG-Fusion architecture exhibits superior performance in modal alignment and semantic fusion.

5 Conclusion

To address the critical need for efficient analysis of blueprint information in the intelligent construction of power grid engineering, this paper presents a multimodal fusion-based joint analysis algorithm for grid blueprint symbols and text. This method achieves precise analysis of power grid blueprints by effectively integrating visual and text features through a dual-stream feature extraction and cross-modal alignment framework. For symbol detection, the YOLOv7 model, augmented with spatial pyramid pooling technology, significantly boosts the detection capability for small-sized electrical symbols. For text recognition, the high-precision PP-OCRv3 engine is employed, with location encoding introduced to enhance spatial perception. Additionally, a symbol-text association matrix is constructed, and its topological connections are modelled using Graph Convolutional Networks (GCNs) to realise semantic associations between symbols and texts. An attention-guided feature fusion module (AG-Fusion) is also designed to dynamically weight and fuse visual and text features, completing the joint analysis task in an end-to-end manner.

To validate the algorithm's effectiveness, systematic experiments were conducted on the self-constructed power grid blueprint dataset GBD-1.0, which includes 217 standard blueprints, 12 types of electrical symbols, and 3862 text instances. The results demonstrate that the average accuracy for symbol detection is 93.7% (mAP @ 0.5), the F1 score for text recognition is 95.4%, and the crucial joint parsing accuracy is 89.2%. These outcomes indicate that the algorithm effectively resolves ambiguity in complex scenarios, such as drawing occlusion and dense text, providing robust technical support for the digital construction of power grids and significantly enhancing the efficiency and accuracy of power grid blueprint information analysis.

In terms of scalability, the current model achieves a training time of 6.7 hours on the GBD-1.0 dataset (single V100 GPU) and an inference speed of 0.18 seconds per graph, meeting engineering application requirements. Future work will explore model pruning and knowledge distillation strategies to further reduce computational overhead. We also plan to validate the model's generalization capability and deployment efficiency on larger datasets such as GBD-2.0.

References

[1] M. Adnan, I. Ahmed and S. Iqbal, "Load flow balancing in super smart grids: A review of technical challenges, possible solutions and future trends from the European prospective," Computers and Electrical Engineering, vol. 117, no., pp. 109265, 2024. https://doi.org/10.1016/j.compeleceng.2024.10926

[2] M. Rizzato, N. Morizet, W. Maréchal and C. Geissler, "Stress testing electrical grids: Generative Adversarial Networks for load scenario generation," Energy and AI, vol. 9, no., pp. 100177, 2022. https://doi.org/10.1016/j.egyai.2022.100177

[3] M. Bernecker, M. Gebhardt, S. B. Amor, M. Wolter and F. Müsgens, "Quantifying the impact of load forecasting accuracy on congestion management in distribution grids," International Journal of Electrical Power & Energy Systems, vol. 168, no, pp. 110713, 2025. https://doi.org/10.1016/j.ijepes.2025.11071

[4] K. Mahmood, S. A. Chaudhry, H. Naqvi, S. Kumari, X. Li and A. K. Sangaiah, "An elliptic curve cryptography based lightweight authentication scheme for smart grid communication," Future Generation Computer Systems, vol. 81, no., pp. 557-565, 2018. https://doi.org/10.1016/j.future.2017.05.002

[5] L. De Langhe, O. De Clercq and V. Hoste, "Position-aware end-to-end cross-document event coreference resolution for Dutch," Natural Language Processing Journal, vol. 13, no., pp. 100184, 2025. https://doi.org/10.1016/j.nlp.2025.100184

[6] K. Li, H. Wu and Y. Dong, "Copyright protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN's international governance regime for AI," Computer Law & Security Review, vol. 55, no., pp. 106056, 2024. https://doi.org/10.1016/j.clsr.2024.106056

[7] A. Peña, A. Morales, J. Fierrez, J. Ortega-Garcia, I. Puente, J. Cordova and G. Cordova, "Continuous document layout analysis: Human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs," Information Fusion, vol. 108, no., pp. 102398, 2024. https://doi.org/10.1016/j.inffus.2024.102398

[8] J. Zhang and Y. Zhang, "DocRouter: Prompt guided vision transformer and Mixture of Experts connector for document understanding," Information Fusion, vol. 122, no., pp. 103206, 2025. https://doi.org/10.1016/j.inffus.2025.103206

[9] Y. Wu, Y. Wu, J. M. Guerrero and J. C. Vasquez, "Decentralized transactive energy community in edge grid with positive buildings and interactive electric vehicles," International Journal of Electrical Power & Energy Systems, vol. 135, no., pp. 107510, 2022. https://doi.org/10.1016/j.ijepes.2021.107510

[10] Y. Wang, W. Lv, W. Wu, G. Xie, B. Lu, C. Wang, C. Zhan and B. Su, "TransTab: A transformer-based approach for table detection and tabular data extraction from scanned document images," Machine Learning with Applications, vol. 20, no., pp. 100665, 2025. https://doi.org/10.1016/j.mlwa.2025.100665

[11] L. Zhu, N. Li and L. Bai, "Embedding-based entity alignment between multi-source temporal knowledge graphs," Engineering Applications of Artificial Intelligence, vol. 133, no., pp. 108451, 2024. https://doi.org/10.1016/j.engappai.2024.108451

[12] R. Du, H. Wang, W. Liu, G. Wang, K. Jiang and H. Ko, "Image Dehazing via RGB-FIR Multimodal Fusion and Collaborative Learning," Pattern Recognition, vol., no., pp. 112206, 2025. https://doi.org/10.1016/j.patcog.2025.112206

[13] P. Schönfelder, A. Aziz, F. Bosché and M. König, "Enriching BIM models with fire safety equipment using keypoint-based symbol detection in escape plans," Automation in Construction, vol. 162, no., pp. 105382, 2024. https://doi.org/10.1016/j.autcon.2024.105382

[14] G. Wang, S. Cheng, A. Du and Q. Zou, "Covariance Attention Guidance Mamba Hashing for cross-modal retrieval," Engineering Applications of Artificial Intelligence, vol. 152, no., pp. 110777, 2025. https://doi.org/10.1016/j.engappai.2025.110777

[15] C. Gerling and S. Lessmann, "Multimodal Document Analytics for Banking Process Automation," Information Fusion, vol. 118, no., pp. 102973, 2025. https://doi.org/10.1016/j.inffus.2025.102973

[16] R. Thanikachalam, A. Muniasamy, A. Alasmari and R. Thavasimuthu, "EffNet-CNN: A Semantic Model for Image Mining & Content-Based Image Retrieval," CMES-Computer Modeling in Engineering and Sciences, vol. 143, no. 2, pp. 1971-2000, 2025. https://doi.org/10.32604/cmes.2025.06306

[17] B. Ju, Y. Liu, J. Liu, P. Sun and L. Song, "CNN-DAG-Editor: A Convolutional Neural Network offloading analyzer with Multi-Objective Dynamic Adaptive Resource Competitive Swarm Optimization," Computer Networks, vol. 268, no., pp. 111374, 2025. https://doi.org/10.1016/j.comnet.2025.111374

[18] S. Andriani, S. Galantucci, A. Iannacone, A. Maci and G. Pirlo, "CNN-AutoMIC: Combining convolutional neural network and autoencoder to learn non-linear features for KNN-based malware image classification," Computers & Security, vol. 156, no., pp. 104507, 2025. https://doi.org/10.1016/j.cose.2025.104507

[19] Q. Jiang, Y. Xiao, G. Zhou, G. Liu, Z. Li, J. Luo, K. He and S. Liao, "Integrating radius margin constraints and class variance for improved CNN-based image recognition," Knowledge-Based Systems, vol. 327, no., pp. https://doi.org/10.1016/j.knosys.2025.11418

[20] A. Muralidharan and S. Mahfuz, "Human Activity Recognition Using Hybrid CNN-RNN Architecture," Procedia Computer Science, vol. 257, no., pp. https://doi.org/10.1016/j.procs.2025.03.04

[21] N. Xiao, "Innovation and Evaluation of Machine Translation Models Combining Reinforcement Learning Algorithms and RNN," Procedia Computer Science, vol. 261, no, pp. 821-828, 2025. https://doi.org/10.1016/j.procs.2025.04.41

[22] A. G. AbdElkader, H. ZainEldin and M. M. Saafan, "Optimizing wind power forecasting with RNN-LSTM models through grid search cross-validation," Sustainable Computing: Informatics and Systems, vol. 45, no., pp. https://doi.org/10.1016/j.suscom.2024.10105

[23] X. Yang, "Research on interactive optimization technology in music education games based on EMD-RNN," Systems and Soft Computing, vol. 7, no., pp. 200305, 2025. https://doi.org/10.1016/j.sasc.2025.200305

[24] M. K. Gupta, A. K. Sharma and S. Goyal, "Adaptive backstepping control for a single-link flexible robot manipulator driven DC motor," Journal of Vibration and Control, vol. 29, no. 15-16, pp. 3445-3457, 2023. https://doi.org/10.1177/10775463221147320

[25] Y.-J. Sun, L.-W. Qiao and S. Ji, "AG-GCN: Vehicle Re-Identification Based on Attention-Guided Graph Convolutional Network," Computers, Materials and Continua, vol. 84, no. 1, pp. 1769-1785, 2025. https://doi.org/10.32604/cmc.2025.062950

[26] D. Meng, J. Zhang, C. Li and Z. Zhao, "Reconsidering the interplay between behaviors: A cross-attentive behavior-aware GCN-based recommendation," Pattern Recognition, vol. 171, no, pp. 112167, 2026. https://doi.org/10.1016/j.patcog.2025.11216

[27] R. Jiao, R. Fan, W. Nan, M. Lu, X. Yang, Z. Zhao, J. Dang, Y. Tian, B. Dong, X. He and X. Luo, "YOLO-MFDNet: An object detection algorithm for multi-scale remote sensing images," Digital Signal Processing, vol. https://doi.org/10.1016/j.dsp.2025.10547

[28] M. Li and N. Yan, "IPD-YOLO: Person detection in infrared images from UAV perspective based on improved YOLO11," Digital Signal Processing, vol. 168, no., pp. 105469, 2026. https://doi.org/10.1016/j.dsp.2025.105469

[29] Q. Zhang, K. Ahmed, M. I. Khan, H. Wang and Y. Qu, "YOLO-FCE: A Feature and Clustering Enhanced Object Detection Model for Species Classification," Pattern Recognition, vol., no., pp. 112218, 2025. https://doi.org/10.1016/j.patcog.2025.112218

[30] D. Zhang, C. Xu, J. Chen, L. Wang and B. Deng, "YOLO-DC: Integrating deformable conversion and contextual fusion for high-performance object detection," Signal Processing: Image Communication, vol. 138, no., pp. 117373, 2025. https://doi.org/10.1016/j.image.2025.11737

[31] H. Liu, Y. Pan and J. Cao, "Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems," Fuzzy Sets and Systems, vol. 467, no. 1, pp. 108584, 2023. https://doi.org/10.1016/j.fss.2022.108584

[32] M. Rahmani, A. Ghanbarpour and P. A. Zarchi, "Output-feedback controller based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities," ISA Transactions, vol. 139, no. 1, pp. 337-348, 2023. https://doi.org/10.1016/j.isatra.2022.12.019

[33] A. Benabdallah, M. Mabrouk and M. Ayadi, "Adaptive backstepping control for a class of uncertain single input single output nonlinear systems," International Journal of Control, Automation and Systems, vol. 20, no. 11, pp. 3651-3660, 2022. https://doi.org/10.1007/s12555-021-0751-4

[34] S. Goyal, A. K. Sharma and R. Sharma, "Nonlinear optimal control for a gas compressor driven by an induction motor," Optimal Control Applications and Methods, vol. 44, no. 4, pp. 1237-1254, 2023. https://doi.org/10.1002/oca.2954

[35] C. Zhang, Y. Liu, J. Li and Z. Wang, "Graph-Based Symbol-Text Association in CAD Diagrams," Pattern Recognition, vol. 131, no. 1, pp. 108917, 2022. https://doi.org/10.1016/j.patcog.2022.108917

[36] Farouk Zouari, K Ben Saad, and M Benrejeb, "Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems," International Review on Modelling and Simulations, vol. 5, no. 5, pp. 2075-2103, 2012.