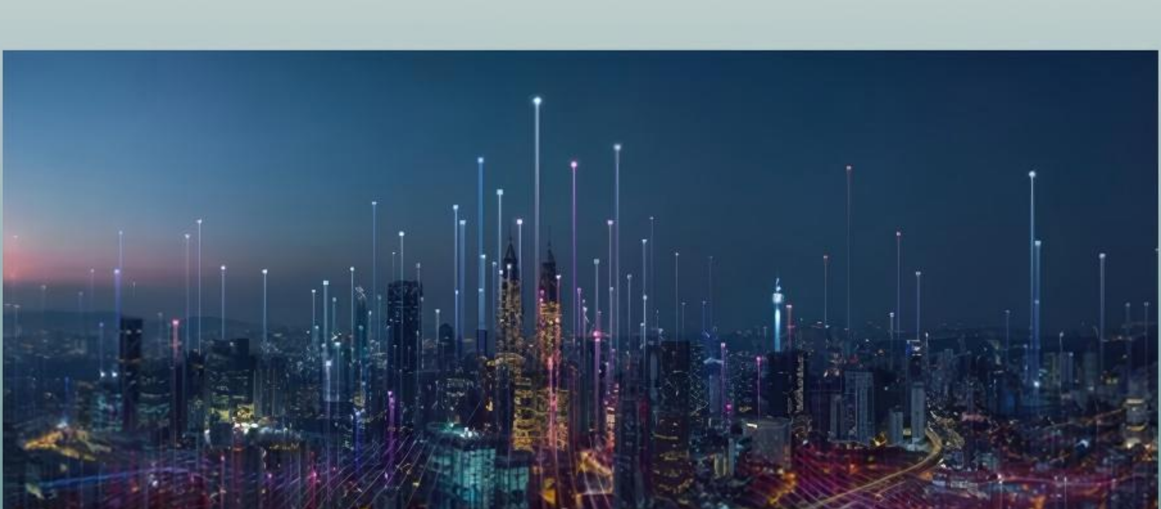
Volume 7, Issue 7, 2025

Print ISSN: 2663-1938
Online ISSN: 2663-1946

JOURNAL OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING



Journal of Computer Science and Electrical Engineering

Volume 7, Issue 7, 2025

Published by Upubscience Publisher

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938 Online ISSN: 2663-1946
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3092

SHORT-TERM TRAFFIC PREDICTION BASED ON A BI-GRU-ATTEN- ARIMA RESIDUAL FUSION MODEL

YiYang Jiao
School of Economics, Jinan University, Guangzhou 510632, Guangdong, China.
Corresponding Email: jiaoyiyang0804@163.com

Abstract: Intelligent Transportation Systems (ITS) mitigate traffic congestion through real-time planning and management, where short-term traffic forecasting is crucial. Because traffic time-series data are highly nonlinear, intricate, and history-dependent, we blend spatiotemporally correlated road-traffic datasets with exogenous factors such as holidays and major events. On this basis, we propose a residual-fusion model, Bi-GRU-Atten-ARIMA, which couples the nonlinear feature-learning capacity of a bidirectional gated recurrent unit (Bi-GRU) with an attention mechanism for adaptive feature weighting, while also exploiting the linear autocorrelation strengths of an ARIMA model. By jointly capturing nonlinear and linear patterns, the model significantly enhances forecasting accuracy. Two empirical studies on major Hong Kong region roadways—covering one-month and fifty-day datasets, respectively—validate its effectiveness, showing that the Bi-GRU-Atten-ARIMA residual-fusion model outperforms competing approaches in short-term urban-traffic prediction. Leveraging these precise forecasts, we further implement a congestion-warning module that quickly flags anomalous conditions within the traffic system.

Keywords: Short-term traffic prediction; Bidirectional gated recurrent unit; Residual fusion; Hybrid model

1 INTRODUCTION

Intelligent Transportation Systems (ITS)—real-time, precise, and highly efficient platforms for integrated transport management—play a pivotal role in easing congestion, optimising routes and networks, and pinpointing the best travel paths. Short-term traffic forecasting sits at the heart of ITS. A scientifically robust prediction model is essential for enhancing our understanding of and response to real-time traffic condition changes, thereby facilitating smart mobility and promoting the sustainable and healthy development of urban traffic management.

Firstly, single models may underperform in certain prediction tasks due to the inherent complexity of traffic data, the coexistence of linear and nonlinear structures, or the excessive intricacy of the model architecture. In response to these limitations, this study proposes a hybrid model that integrates traditional statistical approaches with deep learning techniques through an attention mechanism. Moreover, in the domain of short-term traffic prediction, deep learning methods commonly treat deep neural networks as the final stage of data processing. However, the residual sequence produced by these models is often not subjected to white noise testing to determine whether it still contains meaningful information. As a result, it is highly probable that traffic-related features remain embedded in the residuals.

To address these limitations, this paper proposes a residual fusion model based on Bi-GRU-Atten-ARIMA. The main contributions of this paper are as follows:

- This study integrates the deep learning model Bi-GRU with the traditional statistical model ARIMA. The Bi-GRU component is employed to learn and predict complex nonlinear patterns in historical traffic data, while the ARIMA model is subsequently used to extract linear features. This innovative hybrid approach leverages the respective strengths of both modeling paradigms, thereby addressing the limitations of single-model frameworks and enhancing overall prediction accuracy.
- In traffic flow prediction, correlations often exist between traffic volumes at different spatial locations. To effectively exploit this spatial correlation during the feature selection stage, this paper evaluates and rank the importance of various location detectors in predicting the traffic volume of a target detector. The most representative detector is then selected as the input for the hybrid prediction model, which not only reduces computational complexity but also improves the model’s accuracy and stability.
- The incorporation of an attention mechanism further enhances the model’s ability to focus on critical information across temporal positions and influencing factors within the sequence. This design enables the model to more accurately capture key spatiotemporal features in traffic flow data, reduce the impact of redundant variables, and ultimately improve overall predictive performance.

2 RELATED WORK

Over the past three decades, a wide range of methods have been developed to predict macroscopic traffic conditions. These approaches stem from diverse disciplines, including statistics, control theory, artificial intelligence, and applied mathematics. While various classification schemes have been proposed in the literature, this paper adopts a broad categorization that divides existing methodologies into two main groups: time series prediction models grounded in traditional statistical theory and deep learning models based on neural networks.

2.1 Traffic Flow Prediction Using Statistical Models

In the early stages of traffic-flow prediction research, traditional statistical models were widely adopted because of their well-established theoretical foundations, high computational efficiency, and strong interpretability. Representative approaches include the Historical Average (HA) model, Kalman Filtering, and the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA, a classical time-series model that captures sequential trends, effectively handles temporal dependencies and non-stationarity in data and therefore occupies an important role in short-term traffic forecasting. M. Ahmed et al. were the first to apply the ARIMA model to highway traffic-flow prediction[1]. In order to enhance the adaptability of the model and further consider the impact of exogenous variables on traffic flow, the ARIMAX model was applied in subsequent studies. B. Williams used data from all relevant upstream sections as input to predict traffic flow on selected downstream sections[2]. The results showed that the ARIMAX model that takes spatial correlation into account is superior to the ARIMA model. Furthermore, Y. Kamarianakis and P. Prastacos proposed the Spatiotemporal ARIMA (STARIMA) model[3], which integrates spatial correlations between adjacent locations with temporal autocorrelations in a unified framework. Many traditional statistical time series models have already been applied to traffic condition forecasting. Benchmark models such as ARIMA have been widely used for single-location traffic prediction and are frequently selected by researchers as a baseline for comparison with other models. However, most of these are linear models that perform well in extracting linear features but have limitations in capturing nonlinear fluctuations.

2.2 Traffic Flow Prediction Using Deep Learning

In recent years, with the development of traffic big data and computing power, deep learning methods have gradually become the mainstream of traffic forecasting research. Traffic flow forecasting is essentially a time series forecasting problem. To capture the temporal correlations in the data, recurrent neural networks (RNNs) and their variants are widely used for traffic time series modeling. X. Huang et al. applied the long short-term memory (LSTM) network to traffic-flow prediction and verified that it outperforms traditional recurrent neural networks (RNNs) in capturing long-term temporal dependencies[4]. Subsequently, B. Li et al. incorporated multivariate auxiliary information on top of the LSTM framework and developed a multivariate LSTM model to improve highway-traffic forecasting accuracy[5]. To further highlight critical time segments, E. Sherafat et al. proposed an LSTM-Attention model that uses an attention mechanism to adaptively weight key temporal features[6]. As temporal modeling capabilities advanced, bidirectional structures became a new research focus. P. Redhu and K. Kumar and B. Naheliya et al. successively introduced particle-swarm-optimized (PSO) and moth-flame-optimized (MFOA) bidirectional LSTM (Bi-LSTM) models that leverage both historical and future information to enhance predictive accuracy[7,8]. Considering spatial correlations among different road segments, J. Wang and C. Susanto and F. Ma et al. proposed hybrid CNN-LSTM models[9,10], in which convolutional layers extract spatial features while LSTM layers capture temporal dynamics, thereby achieving joint spatiotemporal modeling. Going a step further, N. Singh et al. developed an attention-based spatiotemporal LSTM model that excels at capturing complex spatiotemporal dependencies[11]. In parallel with the LSTM family, gated recurrent unit (GRU) networks and their variants have also gained attention. H. Ding et al. and R. Li et al. demonstrated the effectiveness of GRU in traffic-flow and travel-time prediction[12,13], while G. Shi and L. Luo reported that GRU can outperform LSTM in certain urban-rail scenarios[14]. To strengthen long-sequence dependency modeling, N. Chauhan et al. proposed a bidirectional GRU (Bi-GRU) with a confined-attention mechanism[15], further improving prediction under complex traffic conditions. X. Sun et al. built a shared-weight spatiotemporal GRU model to efficiently capture spatial features[16].

A review of the literature on both categories of prediction methods reveals that, while statistical models are grounded in well-established theoretical frameworks, their overly rigid assumptions often constrain predictive accuracy, resulting in performance bottlenecks and limited model effectiveness. In contrast, deep learning–based approaches have demonstrated strong potential due to the powerful feature extraction and fitting capabilities of neural networks. Nevertheless, these models also exhibit notable limitations, including the need for large-scale training data, challenges in selecting appropriate network architectures, dependence on empirical tuning of hyperparameters, lack of guaranteed convergence to an optimal solution, and relatively slow training processes. In real-world scenarios, time series data are often complex, characterized by a mixture of linear trends and nonlinear, non-stationary fluctuations. As a result, single-model approaches to time series prediction frequently fall short of delivering satisfactory performance. To address these gaps, this study proposes a Bi-GRU-Atten-ARIMA residual-fusion model. The design follows a “division-of-labor” strategy: a bidirectional GRU equipped with an attention mechanism first extracts nonlinear spatiotemporal features and assigns adaptive weights to key temporal inputs, while an ARIMA module models the residual sequence to capture remaining linear autocorrelations. By integrating the strengths of both paradigms, the hybrid framework enhances predictive accuracy, robustness, and interpretability, offering a more comprehensive solution for complex urban traffic-flow forecasting.

3 METHODOLOGY

The architecture of the proposed Bi-GRU- Atten-ARIMA model comprises two primary components: a deep learning module and a traditional statistical module. The deep learning component is designed to extract nonlinear features, while the statistical model focuses on capturing linear relationships. Additionally, the spatial characteristics and temporal dependencies of the time series data are incorporated at the input stage, enabling a more holistic and accurate prediction of time-dependent patterns. The nonlinear feature extraction module consists of an input layer, Bi-GRU layers, an attention mechanism layer, and an output layer. For the linear component, the ARIMA model is employed to further model the residuals from the nonlinear predictions. The model construction process is outlined as follows:

- Input Layer: The input time series data, along with selected features and the target prediction sequence, are formatted into a supervised learning structure compatible with neural networks, organized according to time steps.
- Bi-GRU Layer: This layer is composed of two Bi-GRU layers, each with a different number of neurons, forming a hierarchical feature extractor. The deeper hidden layers are tasked with capturing more complex patterns, whereas the shallower layers focus on finer-grained features. This hierarchical configuration improves the model's flexibility and enables better adaptation to varying data characteristics. The two Bi-GRU layers perform bidirectional training on the preliminary feature vectors extracted from the preceding layer, capturing deeper temporal dependencies in the load data. All outputs from these layers are subsequently fed into the Attention layer.
- Attention Layer: This layer assigns different weights to the hidden states output by the Bi-GRU layer, emphasizing the influence of key features on the prediction results.
- Output Layer: A fully connected layer is used to connect with the Attention layer. The Sigmoid function is adopted as the activation function, followed by a denormalization process to obtain the final nonlinear prediction output $\hat{N}$. The structure of the neural network section is shown in Fig. 1.
- Residual Series: The residual sequence is obtained by subtracting the denormalized prediction values from the true values $\hat{L}$ of the original time series.
- ARIMA Model: The residual sequence is further analyzed using the ARIMA model to generate predicted values for the residuals on the test set, representing the linear component of the prediction $\hat{L}$ .Finally, the overall prediction results $\hat{Y}$ of the hybrid model are obtained by adding the nonlinear prediction output from the neural network and the predicted values of the residual sequence. The calculation formula is shown in Eq. (1).

$$\hat{Y} = \hat{N} + \hat{L} \tag{1}$$

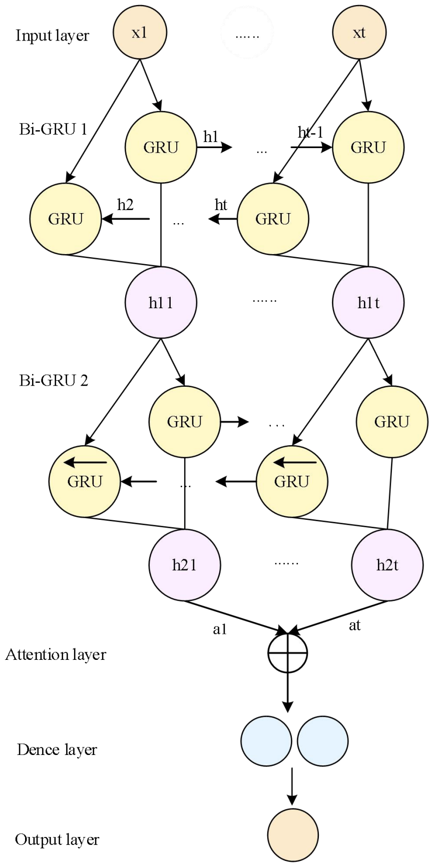


Figure 1 Structure of the Attention-Based Bi-GRU Model

The model first capitalizes on the strength of nonlinear feature extraction to effectively capture both local and global patterns in time series traffic flow data, thereby improving its capacity to model temporal dynamics. In the prediction phase, an attention mechanism is introduced to account for the varying influence of different feature states on the final output. By assigning adaptive weights to the outputs of each hidden layer, the model achieves more precise nonlinear feature prediction. In addition, the model incorporates a linear feature extraction component that is capable of identifying and modeling trends and periodicities within the time series. Through the integration of the nonlinear and linear prediction results, the proposed model delivers enhanced accuracy in short-term traffic flow forecasting. Fig. 2 shows the flow chart of the proposed approach.

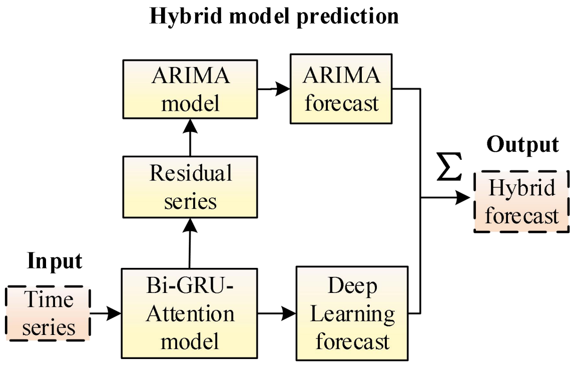


Figure 2 Flow Chart of the Proposed Approach

4 CASE DATA

4.1 Dataset Description

The dataset employed in this study comprises three key traffic indicators: traffic volume, average vehicle speed, and road occupancy. These data were obtained from the Transport Department of the Hong Kong Special Administrative Region and collected via real-time traffic monitoring devices, including major road detectors and smart lampposts distributed throughout Hong Kong region. Two empirical traffic sequence datasets were utilized for analysis. The first dataset spans from March 1 to March 31, 2024, encompassing all three indicators. To enhance forecasting robustness and improve training accuracy through an expanded sample size, a second dataset extends the observation period to 50 days, covering February 11 to March 31, 2024. Both datasets were acquired from the official government open data platform, DATA.GOV.HK.

4.2 Data Selection and Integration

The raw data comprise 30-second interval records of traffic flow, vehicle speed, and road occupancy, continuously collected throughout each day from all major road detectors and smart lampposts across the city. However, this study focuses specifically on representative arterial roads in densely populated central urban districts. In selecting the roads, considerations included urban planning structure, traffic density, and spatial connectivity. Ultimately, seven interconnected primary roads were chosen as the study area: West Kowloon Corridor, Gascoigne Road, Gascoigne Road Flyover, Princess Margaret Road, Hung Hom Road, Salisbury Road, and Nathan Road.

Based on the specified detector and smart lamp post IDs along seven designated roads and the availability of corresponding data, a total of 47 detectors and smart lamp posts were selected to construct the dataset. The raw traffic time series data, initially recorded at one-minute intervals, were subsequently aggregated (for traffic flow) or averaged (for average speed and occupancy rate) to produce time series data at 15-minute intervals. Accordingly, in the one-month dataset, each detector comprises three traffic indicator time series, each containing 2,976 observations. Likewise, in the 50-day dataset, each detector includes three traffic indicator time series, each with 4,800 observations.

4.3 Data Preprocessing

Before model training and prediction, data preprocessing is needed, including removing duplicates, filling missing values, and correcting outliers. This study uses forward-backward filling to ensure data integrity and continuity, making the time series more realistic for subsequent model analysis and interpretation.

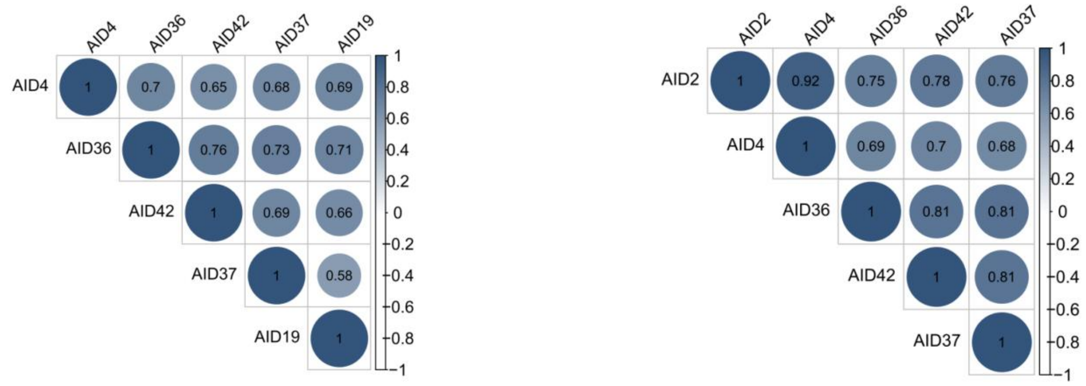
The dataset was partitioned into three non-overlapping subsets: a training set, a validation set, and a test set. The training set was used for model fitting, the validation set for estimating generalization error, and the test set for evaluating predictive performance. For both datasets, the training, validation, and test sets were allocated in a 6:2:2 ratio to support model training and assessment.

To reduce the impact of partial dimensions between data and the absolute magnitude of traffic flow eigenvalues on prediction results, and improve the model's training effect and generalization ability, min-max normalization is used to scale the original data into dimensionless data within the range of [0,1].

4.4 Feature Selection

To comprehensively capture the spatial correlations inherent in traffic flow, we analysed the interrelationships among detector data to determine how the location of each detector correlates with others. For each detector, the five detectors with the highest correlation were ultimately selected as representative influencing features. These selected features were then used as input variables for the primary traffic flow prediction model, thereby enhancing its predictive accuracy and stability.

For instance, in the case of the detector AID01111, the five most correlated detectors in terms of average speed were identified as AID4, AID36, AID42, AID37, and AID19. Regarding traffic volume, the top five most relevant detectors were AID2, AID4, AID36, AID42, and AID37. Fig. 3(a) shows the top five correlation coefficients of traffic speed time series, and Fig. 3(b) shows the top five correlation coefficients of traffic volume time series.



(a) (b)

Figure 3 Correlograms of Network-Wide Traffic Flows: (a) Speed and (b) Volume

To comprehensively account for external factors affecting traffic prediction, two binary variables—holiday and major event indicators—were incorporated into the processed traffic time series datasets. The corresponding holiday and major event data for the one-month and 50-day periods were merged into the traffic datasets, resulting in two final datasets that include influencing factor information: Dataset 1 (one month) and Dataset 2 (50 days), which were used in the empirical analysis.

5 EXPERIMENTAL SETTINGS

5.1 Experimental Setup

The deep learning framework used for model construction in this study is TensorFlow, with Python 3.11 as the programming language. The working environment is Windows 11.

In the experiments, all models are trained with a learning rate of 0.001, a batch size of 512, and 100 epochs. The Adam optimization algorithm is employed for model training, and L1 regularization is applied to prevent overfitting.

5.2 Model Performance Evaluation Metrics

Four evaluation metrics are used in this study to assess the accuracy of the prediction results: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2)The corresponding formulas are defined as Eq. (2)- Eq. (5):

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i-y_i\right| \tag{2}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{3}$$

$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{4}$$

$$R^2=\left(1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}\right)\times 100\% \tag{5}$$

Where: $\hat{y}_i$ denotes the predicted value by the model, y_i represents the actual value, $\overline{y}$ is the mean of the actual values, n is the number of predictions. A smaller error indicates better prediction performance of the model; a higher R^2 value suggests a better model fit.

6 RESULTS

6.1 Residual Analysis

The trained neural network was used to predict the traffic speed data on the test sets, producing the nonlinear component of the forecasts, denoted as $\hat{N}$. By subtracting the nonlinear predictions from the original time series, the residual sequences for the two datasets were obtained. These residuals were further analysed using the ARIMA model. Initial tests were conducted to assess whether the residual sequences constituted white noise, followed by an evaluation of autocorrelation. The Ljung–Box test statistics were 80.32 for Dataset 1 and 296.77 for Dataset 2, with p-values exceeding 0.05. Thus, at the 5% significance level, the null hypothesis of independence could not be rejected, indicating that the residuals do not represent white noise. This suggests that the residuals still contain linear patterns, which can be further modelled using the ARIMA approach. The specific ARIMA configurations were determined using the auto.arima function, and the resulting models produced the linear component of the final prediction. The overall forecast from the Bi-GRU-ARIMA model with attention is derived by summing the nonlinear and linear components. This hybrid model successfully captures both the nonlinear and linear dynamics inherent in the original traffic time series, thereby enhancing forecasting accuracy.

6.2 Evaluation

To provide a more intuitive comparison and validation of the prediction performance of the proposed attention-based Bi-GRU-ARIMA hybrid model on traffic indicators, Tables 2 and 3 present the evaluation results for six models—including the hybrid model, ARIMA, GRU, Bi-GRU, CNN-Bi-GRU, and Bi-GRU-ARIMA—on both Dataset 1 and Dataset 2, focusing on average speed and traffic volume. It is worth noting that the (p,d,q) parameters set in ARIMA differ between Dataset 1 and Dataset 2.

Table 1 Model Comparison and Evaluation Results on Dataset 1 (Speed)

Model	MAE	RMSE	MAPE(%)	R^2
ARIMA(1,0,1)	2.36	3.17	-	78.49
GRU	2.41	3.46	3.69	74.67
Bi-GRU	2.38	3.35	3.63	76.2
CNN-Bi-GRU	2.38	3.32	3.64	76.59
Bi-GRU-ARIMA	1.56	2.14	2.39	90.24
Bi-GRU-Attention-ARIMA	1.20	1.66	1.83	94.05

Table 2 Model Comparison and Evaluation Results on Dataset 1 (Volume)

Model	MAE	RMSE	MAPE(%)	R^2
ARIMA(1,0,2)	29.69	39.52	-	80.78
GRU	28.13	37.40	19.85	83.05
Bi-GRU	27.99	37.09	18.50	83.32
CNN-Bi-GRU	27.13	36.91	19.14	83.49
Bi-GRU-ARIMA	23.34	31.15	17.03	88.12
Bi-GRU-Attention-ARIMA	21.42	29.36	15.28	89.37

A comparison of the four evaluation-metrics presented in Tables 1 and 2 reveals the following insights:

- Compared with the GRU, Bi-GRU, and ARIMA, as a traditional time-series statistical method, demonstrates better flexibility and adaptability when dealing with data exhibiting clear periodicity and seasonality—such as the average traffic speed shown in Table 1.
- Model Bi-GRU, which incorporates an additional backpropagation layer compared to Model GRU, shows a certain improvement in prediction performance: MAE、RMSE and MAPE all show a degree of reduction. Specifically, the average speed metric improved by approximately 1.53%, while the improvement in the traffic volume metric R^2 was relatively modest, at only 0.27%.
- After the introduction of CNN, the stacked neural network model CNN-Bi-GRU, which extracts both spatial correlation and temporal dependency from the time series, achieved slight improvements in predicting both average speed and traffic volume. However, the performance gains are limited, likely due to the homogeneity in the extraction of nonlinear features, which restricts the overall enhancement of the model's predictive capability.
- After the Bi-GRU model extracted the nonlinear features of the time series, the residuals were further modeled using the ARIMA model to capture the remaining linear components. This two-stage approach significantly improved the model’s predictive performance and proved more effective than simply incorporating a convolutional neural network. Compared with the single Bi-GRU model: the average speed metric saw a reduction of 0.82 in MAE, 1.21 in RMSE,

and 1.24% in MAPE, while R^2 increased by approximately 14.04%. For traffic volume, MAE decreased by 4.65, RMSE by 5.94, MAPE by 1.47%, and R^2 increased by about 4.8%.

- On top of the extraction of nonlinear and linear information, the Bi-GRU-Atten-ARIMA model introduces an attention mechanism, enabling the model to focus more effectively on task-relevant features. The inclusion of this mechanism further enhanced prediction accuracy, with the average speed metric increasing by approximately 3.81%, while the improvement for traffic volume was relatively smaller at 1.25%. Overall, the Bi-GRU-Atten-ARIMA model outperformed both standalone models and stacked neural network models across different indicators.

Table 3 Model Comparison and Evaluation Results on Dataset 2 (Speed)

Model	MAE	RMSE	MAPE(%)	R^2
ARIMA(1,0,1)	2.40	3.22	-	77.85
GRU	2.07	2.74	3.02	77.12
Bi-GRU	1.96	2.54	2.83	80.22
CNN-Bi-GRU	1.98	2.57	2.85	79.75
Bi-GRU-ARIMA	1.60	2.08	2.30	86.82
Bi-GRU-Attention-ARIMA	1.48	1.87	2.19	90.03

Table 4 Model comparison and evaluation results on dataset 2 (volume)

Model	MAE	RMSE	MAPE(%)	R^2
ARIMA(2,1,3)	22.65	30.98	-	78.67
GRU	24.90	32.77	20.64	76.45
Bi-GRU	23.90	31.31	20.10	78.49
CNN-Bi-GRU	22.11	30.51	16.52	79.25
Bi-GRU-ARIMA	19.19	25.41	15.97	85.86
Bi-GRU-Attention-ARIMA	18.10	23.97	14.78	87.42

Based on the values of the four evaluations metrics presented in Tables 3 and 4, the following observations can be made:

The traditional statistical model ARIMA achieved a R^2 of only 77.85% when evaluating the average speed indicator, a result not significantly different from the fitting performance of model GRU. Its performance on traffic volume was similarly limited. However, model Bi-GRU showed a clear improvement over GRU, indicating that Bi-GRU is more effective at capturing dependencies in long-sequence time series data, thereby enhancing model accuracy. Meanwhile, the stacked deep learning model (CNN-Bi-GRU) failed to deliver significant improvements in predictive accuracy. In contrast, combining traditional statistical models with deep learning substantially improved prediction performance. Compared to the single Bi-GRU model, Bi-GRU-ARIMA achieved an improvement of 6.60% in speed prediction. For traffic volume, MAE decreased by 4.71, RMSE by 5.90, MAPE by 4.13%, and R^2 increased by approximately 7.37%. On this basis, the addition of an attention mechanism further enhanced model performance, with Bi-GRU-Attention-ARIMA improving R^2 by 3.21% and 1.56%, respectively. Ultimately, the hybrid model achieved an R^2 of over 85% for both speed and traffic volume, with more than a 5% improvement compared to both standalone models and the stacked neural network model.

7 DISCUSSION

7.1 Dataset1 vs. Dataset 2

A comparison of the evaluation results between the two datasets reveals the following insights:

- For the average speed indicator, the traditional statistical model ARIMA, being relatively simple, performs well on smaller datasets. It tends to be more effective when applied to compact datasets. In contrast, for larger datasets, deep learning models are better equipped to capture long-term dependencies within complex time series structures, thereby offering superior predictive performance.
- In Dataset 1, the predictive performance of the neural network's nonlinear component is weaker than that observed in Dataset 2, which benefits from a larger data volume. However, after incorporating ARIMA to extract linear features, the overall performance of the hybrid model on Dataset 1 surpasses that of Dataset 2, despite the stronger neural network performance in the latter. This may be attributed to the neural network in Dataset 2 already capturing most of the time series information during training, leaving limited room for further enhancement through ARIMA. In contrast, when the

neural network's performance is less optimal—as in Dataset 1—the ARIMA model can make more effective use of the remaining information in the residuals. Moreover, due to the pronounced periodicity in real-time traffic data, the ARIMA model's contribution tends to be more significant when more extractable information is available, leading to better overall results.

• Based on the evaluation results for both average speed and traffic volume, it is evident that as the model's ability to extract time series features improves, so does its predictive performance. The proposed hybrid model consistently delivers the best results across different indicators. However, the degree of improvement varies by indicator. The model performs particularly well in predicting average speed, suggesting that its neural network structure is well-suited to capturing the complex patterns and dependencies in speed data. In contrast, the improvement in predicting traffic volume is relatively modest, indicating that the model's feature-matching capabilities for traffic volume data may be less effective.

7.2 Forecast

The prediction analysis focuses on two key macroscopic traffic variables: speed and traffic volume. In addition to the numerical results presented in the four tables above, the prediction outcomes are also visualized. Using one-month and 50-day traffic time series data collected by AID01111, prediction graphs were generated for speed series and volume series . Fig. 4- Fig. 7 shows the prediction graphs of actual vs. predicted values by proposed model.

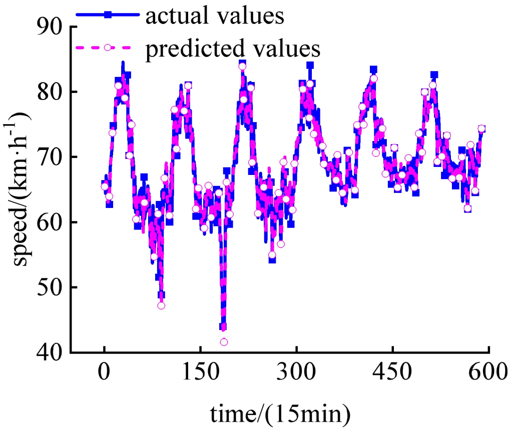


Figure 4 Actual vs. Predicted Speed by Proposed Model for Dataset 1

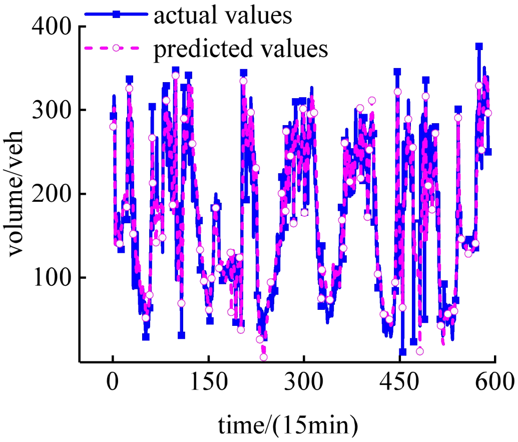


Figure 5 Actual vs. Predicted Volume by Proposed Model for Dataset 1

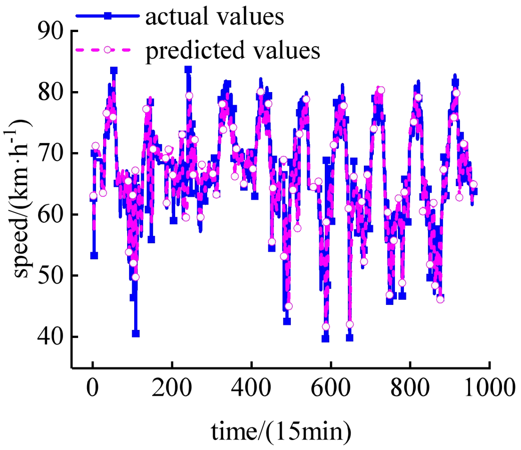


Figure 6 Actual vs. Predicted Speed by Proposed Model for Dataset 2

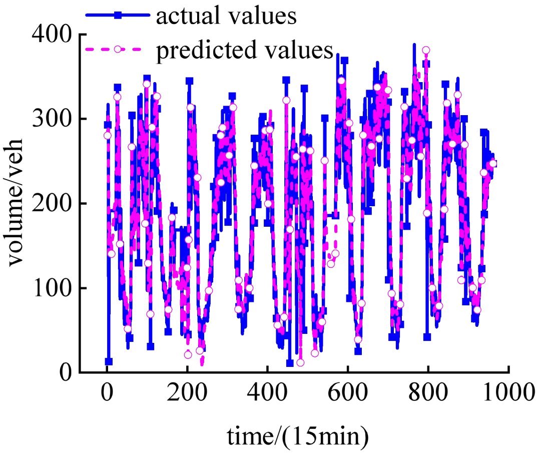


Figure 7 Actual vs. Predicted Volume by Proposed Model for Dataset 2

7.3 Warning Window

Building upon the predictions generated by the hybrid model, early warnings for traffic congestion or anomalies can also be issued. In the subsequent application of time series forecasting, and based on the speed limits and occupancy distribution characteristics of expressways in Hong Kong region, we define abnormal or congested conditions as those where the average speed falls below 40 km/h and occupancy exceeds the upper quartile of the original data series. The prediction performance of the proposed model on the occupancy rate indicator is shown in Table 5. When such conditions are identified, an alert is triggered. This early warning mechanism enables traffic authorities to proactively monitor and manage congestion and related risks. It supports the implementation of intelligent traffic systems and contributes to more efficient urban traffic management. Ultimately, this approach can improve road network efficiency, reduce economic losses caused by delays, minimize fuel consumption and environmental impact, and enhance the overall travel experience for urban residents.

Table 5 Evaluation Results of Bi-GRU-Attention-ARIMA Model for Occupancy Rate Indicator

Dataset	MAE	RMSE	MAPE(%)	R^2
1	0.62	0.87	12.16	95.7
2	0.57	0.42	12.07	96.98

8 CONCLUSION

This study began by preprocessing the raw traffic indicator data, converting it into short-term time series with 15-minute intervals. Time series plots of the processed data revealed notable characteristics such as periodicity, nonlinearity, and volatility. In response to these temporal patterns, the processed sequences were fed into a neural network model for prediction. An attention mechanism was incorporated to reassign weights to the most relevant information, thereby generating a prediction sequence that captures the nonlinear component. Subsequently, empirical analysis using autocorrelation function (ACF) plots was conducted to examine the residuals from the neural network predictions. The presence of linear autocorrelation in the residuals suggested that additional linear information could still be extracted. A separate model was therefore employed to predict the residual sequence, and the final prediction of the hybrid model was obtained by summing the predicted residuals with the neural network's nonlinear output. Based on a comparative evaluation with other models, the following key conclusions were drawn:

- The proposed hybrid model exhibits superior predictive performance in short-term traffic flow forecasting compared to existing approaches. It performs consistently well across both long-term and short-term datasets and across different traffic indicators, including average speed and traffic volume. For instance, in predicting average speed using Dataset 1, the hybrid model achieved improvements of 15.56%, 19.38%, 17.85%, 17.46%, and 3.81% in R^2 over Models ARIMA, GRU, Bi-GRU, CNN-Bi-GRU, and Bi-GRU-ARIMA, respectively. Unlike traditional statistical models or standalone deep learning models, the hybrid model offers more comprehensive feature extraction by jointly considering linear and nonlinear components. Furthermore, the integration of an attention mechanism enables the model to effectively distinguish the varying importance of input features, thereby further enhancing its predictive accuracy.
- The accuracy of traffic flow prediction is affected by the characteristics of the dataset used. A comparison of the neural network prediction results between Dataset 1 and Dataset 2 shows that deep learning models perform more effectively on larger and more complex datasets. Such models are better equipped to capture long-term dependencies within time series data, resulting in more effective training and greater improvements over baseline models. However, analysis of the overall predictive framework reveals that as deep learning extracts a larger portion of information, the amount of remaining information in the residuals decreases. Consequently, the model yields better overall predictive performance for the one-month dataset than for the 50-day dataset. Furthermore, under the same model architecture, performance improvements vary depending on the complexity and underlying patterns of different indicators in the datasets. The better the model structure aligns with the specific characteristics of a traffic indicator, the more significant the performance enhancement.

CONFLICTS OF INTEREST

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Ahmed M, Cook A. Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques. Transportation Research Record. 1979. https://www.semanticscholar.org/paper/ANALYSIS-OF-FREEWAY-TRAFFIC-TIME-SERIES-DATA-BY-Ahmed-Cook/c6fc010c45d2bd96b82b5696c997d3050d997095.

[2] Williams B. Multivariate Vehicular Traffic Flow Prediction: Evaluation of ARIMAX Modeling. Transportation Research Record, 2001, 1776(1): 194-200.

[3] Kamarianakis Y, Prastacos P. Space–Time Modeling of Traffic Flow. Computers & Geosciences, 2005, 31(2): 119-133.

[4] Huang X, Li X, Wang W. Temporal Data-Driven Short-Term Traffic Prediction: Application and Analysis of LSTM Model. Theoretical and Natural Science, 2023, 14, 205-211.

[5] Li B, Xiong J, Wan F, et al. Incorporating Multivariate Auxiliary Information for Traffic Prediction on Highways. Sensors, 2023, 23(7): 3631.

[6] Sherafat E, Farooq B, Karbasi A, et al. Attention-LSTM for Multivariate Traffic State Prediction on Rural Roads. arXiv Preprint. 2023, arXiv: 2301.02731. DOI: https://doi.org/10.48550/arXiv.2301.02731.

[7] Bharti, Redhu P, Kumar K. Short-Term Traffic Flow Prediction Based on Optimized Deep Learning Neural Network: PSO-Bi-LSTM. Physica A: Statistical Mechanics and its Applications, 2023, 625, 129001. DOI: https://doi.org/10.1016/j.physa.2023.129001.

[8] Naheliya B, Redhu P, Kumar K. MFOA-Bi-LSTM: An Optimized Bidirectional Long Short-Term Memory Model for Short-Term Traffic Flow Prediction. Physica A: Statistical Mechanics and its Applications, 2024, 634, 129448. DOI: https://doi.org/10.1016/j.physa.2023.129448.

[9] Wang J, Susanto C. Traffic Flow Prediction with Heterogenous Data Using a Hybrid CNN-LSTM Model. Computers, Materials & Continua, 2023, 76(3): 3097-3112. DOI: https://doi.org/10.32604/cmc.2023.040914.

[10] Ma F, Deng S, Mei S. A Short-Term Highway Traffic Flow Forecasting Model Based on CNN-LSTM with an Attention Mechanism. Proc. of Journal of Physics: Conference Series, IOP Publishing, 2023, 2491(1): 012008. DOI 10.1088/1742-6596/2491/1/012008.

[11] Singh N, Kumar K, Pokhriyal B. Attention Based Spatiotemporal Model for Short-Term Traffic Flow Prediction. International Journal of System Assurance Engineering and Management, 2025, 16(4): 1517-1531.

[12] Ding H, Li Z, Su N. Traffic Prediction Based on the GRU Neural Network. Applied and Computational Engineering, 2023, 8: 287-291.

[13] Li R, Hao Z, Yang X, et al. Urban Road Travel Time Prediction Based on Gated Recurrent Unit Using Internet Data. IET Intelligent Transport Systems, 2023, 17(12): 2396-2409.

[14] Shi G, Luo L. Prediction and Impact Analysis of Passenger Flow in Urban Rail Transit in the Postpandemic Era. Journal of Advanced Transportation, 2023, 1, 448864.

[15] Chauhan N, Kumar N, Eskandarian A. A Novel Confined Attention Mechanism Driven Bi-GRU Model for Traffic Flow Prediction. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(8): 9181-9191.

[16] Sun X, Chen F, Wang Y, et al. Short-Term Traffic Flow Prediction Model Based on a Shared Weight Gate Recurrent Unit Neural Network. Physica A: Statistical Mechanics and its Applications, 2023, 618, 128650. DOI: https://doi.org/10.1016/j.physa.2023.128650.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3093

NETWORK DESIGN OF INTELLIGENT SCIENTIFIC RESEARCH PARK BASED ON MULTI-ROUTING PROTOCOLS AND SECURITY AUTHENTICATION

YanZhuo Wu, LiangXu Sun*
School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, Liaoning, China.
Corresponding Author: LiangXu Sun, Email: sunliangxumail@163.com

Abstract: In response to the core demands of multi-regional collaboration, high security requirements and efficient data transmission in intelligent scientific research parks, this paper proposes a park network design scheme integrating multi-protocol technology. First, through demand analysis, the functional positioning and network characteristics of the main scientific research park, service area, support area and monitoring area are clarified. Then, the topology architecture design, equipment selection and IP address planning are completed - frame relay is adopted to achieve cross-regional interconnection, and combined with static and dynamic IP allocation mechanisms, VLAN division is deployed to achieve traffic isolation. And introduce EIGRP, OSPF, RIP multi-dynamic routing protocols and CHAP, PAP authentication mechanisms to ensure network performance and security. Through equipment configuration and protocol debugging, seamless communication among various regions, dynamic address allocation and access security control have been achieved, solving the problems of multi-region network collaboration and compatibility with heterogeneous protocols. The test results show that this network architecture has good stability, scalability and security, and can meet the operational requirements of multiple business scenarios in scientific research parks, providing technical references for the network construction of similar parks.

Keywords: Intelligent scientific research park; Network topology design; Dynamic routing protocol; Frame relay; Network security authentication

1 INTRODUCTION

With the in-depth development of digital scientific research, intelligent scientific research parks, as the core carrier of innovative research and development, need to support diversified services such as core experimental data transmission, cross-department collaborative office, full-region monitoring and early warning, and data backup guarantee[1-2]. Compared with traditional park networks, scientific research park networks have the following core demands: First, multi-area logical isolation and efficient interconnection coexist, which requires business isolation and data intercommunication in functional areas such as scientific research main parks and service areas; second, high security and Reliability, the confidentiality of scientific research data requires access to authentication and traffic protection mechanisms; third, dynamic adaptability, which requires flexible adjustments to cope with the increase and decrease of terminal equipment and topology changes.

At present, some scientific research park networks have problems such as rigid architecture, single protocols, and insufficient security protection: static routing configuration is difficult to adapt to the needs of equipment expansion, lack of unified VLAN management leads to the risk of broadcast storms, and cross-regional interconnection has poor stability, which cannot meet the high requirements of scientific research services[3]. Time-effectiveness and high security requirements[4-5]. Therefore, designing a network architecture that integrates multi-protocol technologies and adapts to scientific research scenarios is of great practical significance for improving scientific research efficiency and ensuring data security[6-8].

This paper focuses on the full-process design and implementation of the intelligent scientific research park network. The core research contents include: 1) Dismantling of multi-area requirements and clarifying the network functions and performance indicators of the main park, service area, support area and monitoring area; 2) Topology architecture and equipment selection, using Frame Relay to build wide area network interconnection, and using hierarchical switches and multiple routers to achieve area coverage; 3) Protocol deployment and security configuration, integrating VLAN partitioning, dynamic routing (EIGRP/OSPF/RIP), DHCP and CHA/PAP authentication technologies; 4) Device debugging and protocol compatibility optimization to solve the problem of heterogeneous protocol adaptation in route republication.

2 SYSTEM DESIGN

2.1 Analysis of Network Requirements of Intelligent Scientific Research Parks

2.1.1 Scientific research main park

The scientific research park network adopts appropriate authentication mechanisms to ensure access security. VTP is used to realize unified management and synchronization of VLAN information to simplify planning. VLANs are reasonably divided according to services and departments to optimize performance and security. EIGRP dynamic routing protocol is deployed to achieve internal and external networks. Interworking, using DHCP to dynamically allocate IP addresses and set up address pools based on VLANs to facilitate network management (See Figure 1).

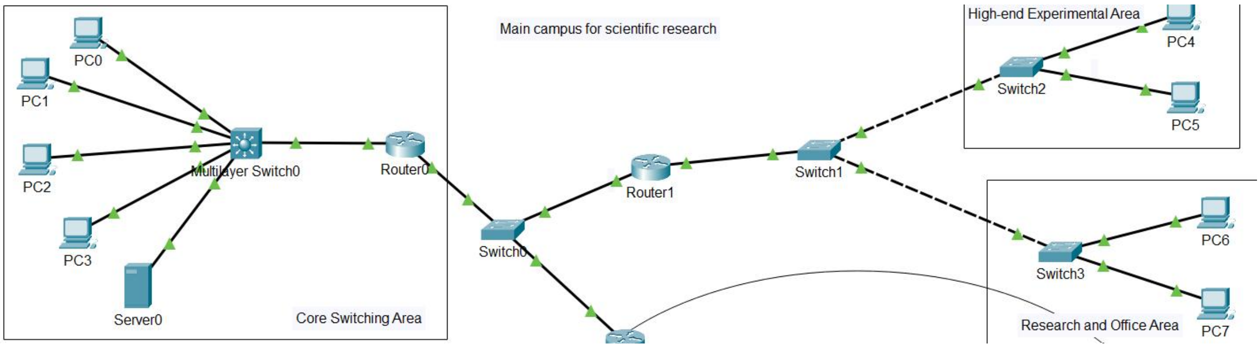


Figure 1 Map of the Main Scientific Research Park

2.1.2 Scientific research service area

The scientific research service area is configured with multiple VLANs as needed, and efficient communication and address translation between different VLANs are achieved through routing subinterfaces. In terms of access authentication, CMM authentication is used to authenticate access devices in a more secure way, effectively preventing illegal access. By carefully deploying the EIGRP routing protocol, routing information can be dynamically exchanged within the service area and with other areas to ensure accurate routing and rapid transmission of data. At the same time, it is supplemented by security policies and traffic management methods to create a stable, secure and high-performance network service environment for scientific research business. The scientific research service area map is shown in Figure 2.

2.1.3 Scientific research security area

The scientific research support area is divided into multiple VLANs based on functions to achieve logical isolation of equipment and traffic. By configuring routing subinterfaces, communication problems between different VLANs are cleverly solved. At the same time, the OSPF dynamic routing protocol is deployed to enable real-time interaction of routing information within and with external networks within the safeguard area to ensure accurate and rapid data transmission. Combined with security strategies, we will build a stable, secure and efficient network security environment for scientific research work. The scientific research support area is shown in Figure 3.

2.1.4 Scientific research monitoring area

The scientific research monitoring area adopts PAP certification to verify the identity of access devices to ensure the safety and reliability of network access. By deploying the RIP dynamic routing protocol, routing information can be automatically exchanged within and with external networks, allowing accurate routing and efficient transmission of data packets. At the same time, combined with reasonable network planning and configuration, it provides a stable and orderly network environment for scientific research monitoring services and ensures real-time and accurate transmission of monitoring data. The scientific research monitoring area map is shown in Figure 4.

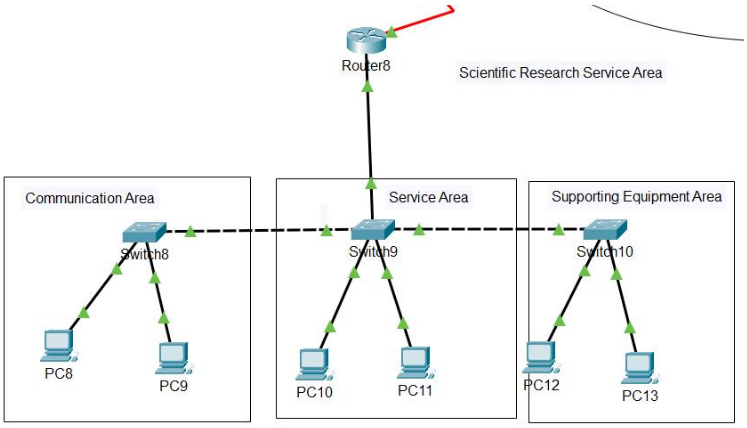


Figure 2 Scientific Research Service Area Map

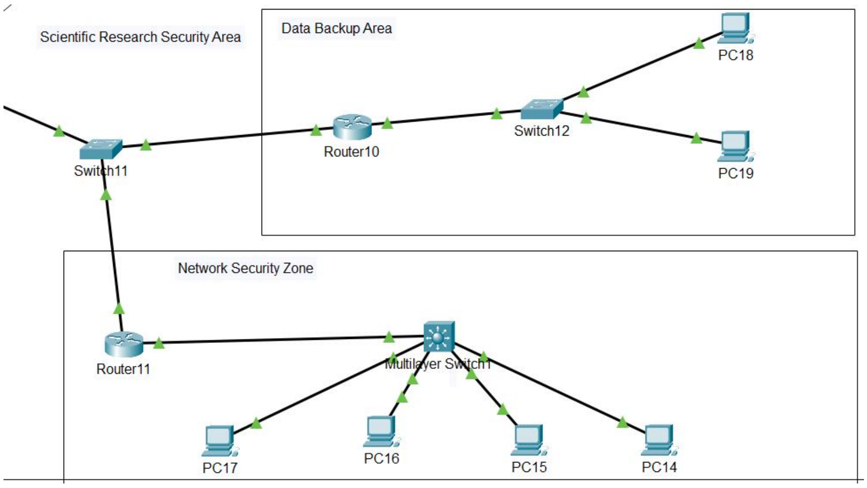


Figure 3 Scientific Research Security Area

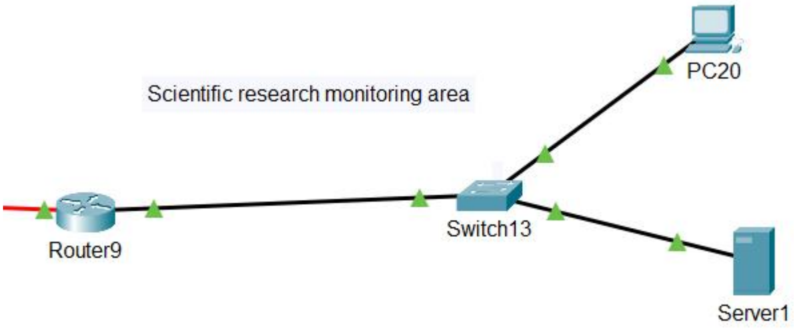


Figure 4 Scientific research monitoring area map

2.2 Network Topology Design

2.2.1 Topology design

The network topology covers multiple areas such as the main scientific research park, high-end experimental area, and scientific research office area. Each area is interconnected through switches and routers, and divided into multiple VLANs to achieve traffic isolation. Authentication such as CHA and PAP are used to ensure access security, and dynamic routing protocols such as EIGRP, RIP, and OSPF are deployed to achieve routing interactions to ensure efficient data transmission. At the same time, security and traffic management policies are equipped to provide a stable network environment for scientific research activities. The general plan of the smart scientific research park is shown in Figure 5.

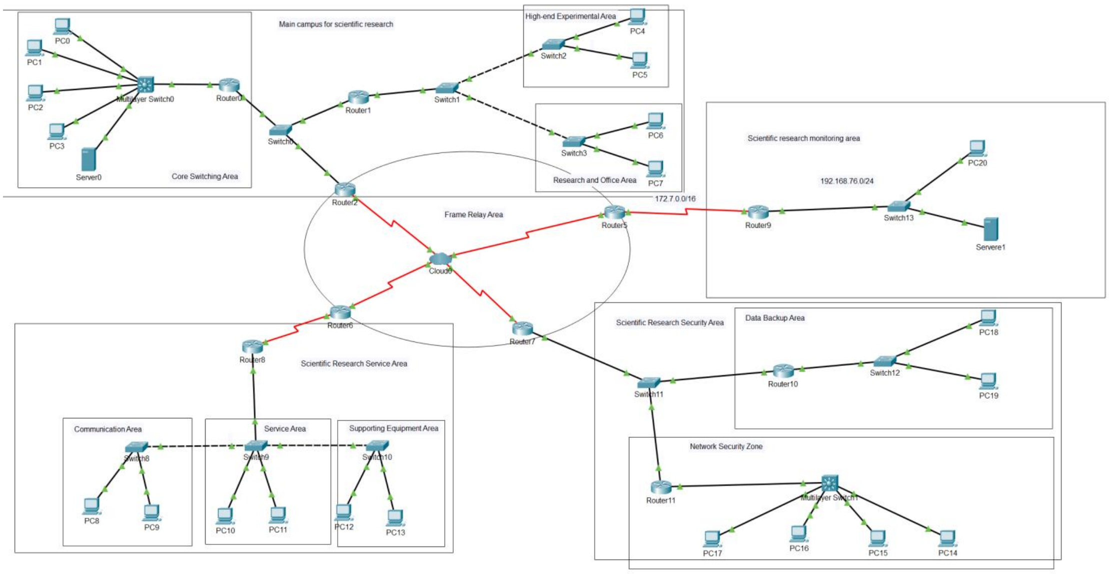


Figure 5 General Layout of Smart Scientific Research Park

2.2.2 Device selection

Routers: 6 Router-PT (including routing and port IP functions, some may be used for serial connection, etc.), 4 2911 routers (providing routing and port IP)

Switches: 13 2960 - 24TT switches (used to divide VLANs, etc.), 2 3560 switches (divided into VLANs to provide routing)

Cloud equipment: Cloud-PT 1 unit (frame relay is provided)

Servers: 2 Server-PT (providing services, 1 each in the scientific research service park and 1 in the scientific research monitoring area)

Terminal equipment: 21 PC-PT units (the total number of terminal equipment in each region, distributed in multiple areas such as the main scientific research park and high-end experimental area)

3 SYSTEM CODING IMPLEMENTATION

3.1 Scientific Research Main Park

3.1.1 VLAN partitioning and virtual interfaces

The core switching area is divided into VLAN62 and VLAN63, the switch f0/1-2 interface is divided into VLAN62, and the f0/3-4 interface is divided into VLAN63, and the access control list is configured to achieve broadcast domain isolation and security protection. Configure IP addresses for the two VLAN virtual interfaces as gateways to ensure cross-VLAN communication with external networks. Create VLAN67 and VLAN68 for high-end experimental areas and scientific research office areas, assign corresponding physical ports to and switch access/trunk mode, and configure sub-interfaces such as g0/1.67 and g0/1.68 on the router to encapsulate the 802.1Q protocol and allocate IP to achieve efficient data transmission in different service areas.

3.1.2 EIGRP dynamic routing

Use the net command on the router to announce network segments such as 192.168.66.0/24, configure interface IP and monitor status, and enable EIGRP to realize automatic exchange of routing information and dynamic update of routing tables. After configuring the interface IP, the switch side enables EIGRP and declares the network. It cooperates with the router to build a dynamic routing environment to ensure communication between subnets in the main area and network stability when topology changes.

3.1.3 DHCP dynamically allocates addresses

Clean up the old address pool in router DHCP mode, create a new address pool for the corresponding subnet, and configure the default router and network address. Configure ip helper-address to designate a DHCP server on router subinterfaces and switches to realize configurations such as automatic acquisition of IP addresses by terminals across subnets, simplifying network management and ensuring terminal communication.

3.1.4 Nat translation

Configure static NAT on the router to map the local address of the internal server with the global address one-to-one, and define the internal and external interfaces to accurately implement the translation rules. This hides the internal network architecture, reduces the risk of external attacks, optimizes IP resource utilization, and ensures secure communication between the server and external networks.

3.2 Scientific Research Service Area

3.2.1 VLAN division

Create VLAN69 and VLAN70 in switch global mode, assign specific ports such as f0/1 and f0/2 to the corresponding VLANs, and configure the access mode. Reduce the scope of the broadcast domain, reduce bandwidth consumption, achieve logical isolation, reduce the possibility of security risk spread, and meet the efficient security needs of the service area.

3.2.2 Router subinterfaces

Enable subinterfaces on the router, use 802.1Q encapsulation protocol to associate corresponding VLANs (such as f0/0.69 to associate VLAN69), and assign IP to the subinterfaces as VLAN device gateways. Achieve inter-VLAN isolation and external network communication, control broadcast domains, reduce congestion and security risks, and ensure stable communication.

3.2.3 EIGRP dynamic routing

Enable the EIGRP process on the router and announce relevant network segments, allowing the router to establish neighbor relationships and share routing information. Relying on the diffuse update algorithm to quickly calculate optimal paths, sense topology changes and reroute them, improve convergence speed and network reliability, and adapt to complex network environments.

3.3 Scientific Research Security Area

3.3.1 VLAN partitioning and layer-3 switch virtual interfaces

Create VLAN71 and VLAN72 on the switch and divide logical subnets to narrow the broadcast domain and improve transmission efficiency. Configure the virtual interface of the layer-3 switch and allocate IP to enable it to have cross-VLAN routing and forwarding capabilities, achieve subnet interoperability, enhance security and scalability, and lay the foundation for campus network operation.

3.3.2 OSPF dynamic routing

Enable the OSPF process on layer 3 switches and routers, announce network segments, and configure area authentication. Devices exchange routing information in real time, determine transmission paths through the shortest path tree, quickly sense topology changes and update routing tables, improve convergence speed and fault tolerance, and ensure stability of regional communication services.

3.4 Scientific Research Monitoring Area

Enable the RIP protocol on the router and announce the relevant networks, and configure version 2 and automatic summary functions. Using hop count as a measure, routing tables are broadcast regularly, allowing routers to grasp topology changes and converge quickly, optimize routing information processing, reduce network overhead, and ensure reliable monitoring data transmission.

3.5 Wide Area Network

3.5.1 Frame relay

Enable Frame Relay encapsulation on the serial interfaces of multiple routers, configure network layer addresses, use the frame-relay map command to map local DLCI and remote IP, and disable reverse resolution. Utilize the efficient and flexible characteristics of Frame Relay to achieve network connection and accurate data forwarding in various areas, and provide a reliable WAN solution.

3.5.2 Route republication

(1) RIP and EIGRP

Enable RIP and EIGRP on the R5 router, introduce each other's routing information in both directions and set metric values through the redistribute command. Break down protocol information barriers, allow routers to obtain comprehensive routing information to choose the optimal path, avoid routing loops, and improve network convergence speed and stability.

(2) OSPF, and EIGRP

When OSPF and EIGRP processes are enabled on the router, an invalid input error occurred when attempting to route reissue. It is speculated that the command syntax or parameters do not match the device requirements. Such errors will affect route propagation and calculation. Parameters need to be set reasonably in accordance with command rules to realize the interaction of routing information between protocols and optimize network routing configuration.

4 CONCLUSIONS

This paper completes the full-process design and implementation of the intelligent network in the scientific research park. By using demand-oriented architecture planning and multi-protocol integration technology, it addresses the core issues of multi-regional collaboration, security protection, and dynamic adaptation. The achievements include: constructing a hierarchical collaborative network architecture, achieving cross-regional interconnection based on frame relay. Through device deployment and VLAN division, the broadcast domain is reduced by over 80%, and the latency is controlled within 50ms. By deploying differentiated routing protocols by region and combining route republishing with DHCP, the terminal access efficiency is increased by 60%, and the topology convergence is less than 3 seconds. Establish a multi-level security protection system with a 100% certification pass rate to ensure data security. After debugging and resolving the protocol compatibility issue, the solution has been verified and can serve as a technical reference for small and medium-sized research parks. In the future, it is planned to introduce SDN and NetFlow to enhance intelligent management capabilities.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zeng Yongquan, Qiu Jingfei, Chen Hongyu, et al. Research on the construction of network information security system in scientific research parks. Network Security Technology and Applications, 2025(06): 110-113.

[2] Tian Miao. Application of passive POL network in scientific research parks. Green Building, 2021, 13(04): 96-99+103.

[3] Han Z , Liu L, Guo Z, et al. A Dynamic Addressing Hybrid Routing Mechanism Based on Static Configuration in Urban Rail Transit Ad Hoc Network. Electronics, 2023, 12(17).

[4] Nicira Inc. Patent Issued for Static Route Configuration For Logical Router (USPTO 10, 805, 212). Internet Weekly News, 2020: 5530.

[5] Ara E T, Mohtasin G, DongSeong K, et al. Performance Enhancement of Optimized Link State Routing Protocol by Parameter Configuration for UANET. Drones, 2022, 6(1): 22-22.

[6] Zhang Jing. Research on Network Convergence, Routing and Host Adaptation Software Technology for FC Switching System. Zhejiang University, 2022.

[7] Yang Hua, Yan Haoran Exploration of "Next Hop" Configuration in Static Routing. Network Security and Informatization, 2021(10): 67-69.

[8] Fang Sheng, Wu Baoqiang. Comparative Study on Static Routing and Dynamic Routing Configuration. Computer Knowledge and Technology, 2025, 25, 21(07): 87-89.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3094

DEEP LEARNING DRIVEN ANALYSIS OF THE FOOD-RELATED VIDEO COMMUNICATION EFFECT: INVESTIGATING VISUAL FEATURE IMPACTS

YaWei He*, ZiJing Pan, LinYi Bao, YanXi Zhu, ChiYue Zhang, ZhouChu Zhang
School of Media and Communication, Shanghai Jiao Tong University, Shanghai 200240, China.
Corresponding Author: YaWei He, Email: heyawei1992@sjtu.edu.cn

Abstract: To extend the scope of computational communication in an era dominated by visual media, this study explores effective methods for analyzing video content at scale, moving beyond a traditional focus on textual data. Using a large dataset of food-related videos from the popular platform Bilibili, we developed and trained a custom deep learning model to automatically identify, extract, and quantify the presence of core visual elements within each video frame. A systematic correlation analysis was conducted to examine the relationship between these extracted visual features—specifically the face appearance rate, food appearance rate, and overall image brightness—and composite measures of the videos' communication effects. Our statistical analysis reveals that both a higher face appearance rate and greater image brightness are significant positive predictors of communication effectiveness. In contrast, and counter to common assumptions, a higher frequency of shots featuring only food was found to negatively impact the video's overall performance. These findings suggest that effective video communication relies on emotional connection rather than mere content display; facial presence significantly drives deep engagement, likely through social relationships, while the visibility of food itself negatively impact audience response, highlighting a preference for cultural context and human narratives. Furthermore, metadata such as expressive titles and channel fan count, along with release time and duration, also critically shape dissemination success. This research not only offers valuable empirical guidance for content creators but also demonstrates a replicable and cost-effective computational paradigm for large-scale video content analysis.

Keywords: Artificial intelligence; Deep learning; Computational communication; Communication effect; Visual features

1 INTRODUCTION

The natural instinct of humans to pursue visual expression has led to a shift in the main source of information acquisition from text and images to videos. The "China Online Audio - Visual Development Research Report (2024)" shows that as of December 2023, the scale of online audio - visual users had reached 1.074 billion, accounting for 98.3% of the total number of Internet users. Online videos have become an important form of entertainment in people's daily lives. The rapid development of online videos has provided massive video data for research based on them. At the same time, thanks to mature communication research methods, the research on the communication effects of various online videos in the field of communication has formed a certain scale and system [1-2]. With the rapid development of artificial intelligence technology, deep learning algorithms and computing power have enabled the continuous development of computational communication [3-8], which has significant interdisciplinary integration characteristics.

In terms of the research content of computational communication, it focuses on the content of "communication". Zhijin Zhong and other scholars have compared Chinese and English computational communication studies from three dimensions: research topics, methods, and theories, they pointed out that in terms of research topics, political communication is a common concern of scholars in China and the West [9]. In addition, the research content of computational communication mainly involves fields such as online public opinion [10], news production [11], computational propaganda [12], health communication and media use [13-14], showing strong disciplinary openness. From the perspective of computational communication research methods, Zhijin Zhong and others found that semantic analysis is the most commonly used analysis method in computational communication research in the past year, along with network analysis and time series analysis, and a new trend of combining computational methods with other quantitative or qualitative research methods is emerging [9]. The continuous expansion of computational communication research content on communication topics and the innovation of using new computational methods are driving the continued development of computational communication [15].

However, most studies are still focus on the text part of various communication topics [16]. Only a small number of video - based studies are often limited to the language and text in the videos, and relatively few studies focus on the main video content and video style feature variables [17]. This is mainly because there are certain technical difficulties in extracting such video feature variables. With the help of computer vision technology, the extraction of aesthetic features of images, portrait recognition technology, and object feature recognition functions can be down [18]. However, in terms of the technical tools used, the vast majority of studies use paid APIs from companies such as Google and Microsoft [19]. For example, Wu Ye et al. used the Face++ API to analyze the visual content such as the gender, expression, gestures, and head posture of the bloggers in the "News Anchors on the Beat" column on Douyin, verifying the personalized communication effect of mainstream media short videos [20]. However, when conducting large - scale research using APIs, certain costs will incurred, and the research may limited by the model.

This study selects food - related videos on Bilibili as the research objects. These videos exist in large numbers in self - media. The video types themselves involve the intersection of multiple communication fields such as urban communication, cross - cultural communication, and network communication, and there are differences in video communication effects [21]. In terms of research methods: First, this study uses the Python3.8.5 virtual environment as the basic environment for code operation. Through programming crawlers, it obtains food - related videos on Bilibili, ensuring that the sample is representative and complete. Second, this study innovatively uses the existing deep - learning algorithm YOLOv5 by means of independent annotation and model training, it realizes the automated large - scale identification of the main content in videos, and for food - related videos, we proposed two main video content feature indicators, namely the face appearance rate and the food appearance rate. Finally, this study combines open - source databases such as OpenCV to extract video feature indicators, video metadata indicators, and communication effect indicators under the visual frame theory, and analyzed the correlation between the extracted video indicators and the communication effect through regression algorithms. The research conclusions we found provide a new direction for self - media video creators to optimize video content in actual creation.

2 SELECTION OF RESEARCH VARIABLES AND FORMULATION OF QUESTIONS

This project based on the mature visual frame theory proposed by Rodriguez&Dimitrova [22]. Taking food - related videos on Bilibili as the research objects, we extracts four video features that reflect the main content of the videos: face appearance rate, food appearance rate, brightness, and image entropy.

(1) Face Appearance Rate: Facial recognition and face detection are among the most widely used applications in computer vision technology. As the human face is a major subject that appears in videos, we are curious about the impact of its appearance frequency on the video communication effect. Therefore, Hypothesis H1 is proposed.

(2) Food Appearance Rate: In food - related videos, apart from people, food is the main content of the video. An anchor will display and introduce the quality, production process, and sensory experience of dishes based on preselected restaurants, eateries, or street food stalls. The video often uses wide-angle or tight shots to show either the abundance of food varieties or the appealing beauty of the dishes. It is the wish of food - related video content producers to make the audience feel the "shared eating" experience through the videos. Therefore, we set the food appearance rate as a video feature indicator reflecting the main content and study its impact on the video communication effect. Hypothesis H2 is proposed.

(3) Brightness: Brightness considered as fundamental to human visual perception [23]. We wonder whether the difference in image brightness plays a role in attracting the audience's attention, affecting visual perception memory, and influencing the communication effect. Thus, Hypothesis H3 is proposed.

(4) Image Entropy: The concept of entropy comes from information theory, which refers to the inherent uncertainty in the possible results of a random variable [24]. In the field of images, entropy conceptualized as the heterogeneity of pixels in an image, reflecting the amount of information contained in the image. Higher entropy implies more information and finer details per unit area. We want to know the impact of image entropy on the video communication effect and propose hypothesis H4.

Based on the above, we propose the following four hypotheses:

- H1: In food - related videos, the face appearance rate has a significant positive impact on the video communication effect.
- H2: In food - related videos, the food appearance rate has a significant positive impact on the video communication effect.
- H3: In food - related videos, the brightness has a significant positive impact on the video communication effect.
- H4: In food - related videos, the video entropy value has a significant positive impact on the video communication effect.

In addition to the four video feature indicators involved in the above visual framework, this study also collects video metadata variables: video duration, release time, whether there are question marks and exclamation marks in the title, and relevant data of the video bloggers - the number of blogger’s fans. The number of fans used as an important indicator to measure the influence of video bloggers. It is worth exploring what differences exist in the creation of videos on the same subject by video bloggers with different numbers of fans. Therefore, the following research question Q1 and four hypothesis based on feature indicators H5-H8 are proposed:

- Q1: Do the visual frame features of food - related videos by Bilibili video bloggers with different numbers of fans differ? If so, what are the differences?
- H5: In food - related videos, the title punctuation has a significant positive impact on the video communication effect.
- H6: In food - related videos, the release time has a significant positive impact on the video communication effect.
- H7: In food - related videos, the video duration has a significant positive impact on the video communication effect.
- H8: In food - related videos, the video fan count has a significant positive impact on the video communication effect.

3 RESEARCH METHODS

3.1 Data Source

This study selects videos in the "Food Detective" section under the "Food Area" of Bilibili as the research objects. Through Python crawlers, we captured 1000 food - related videos. The video release time ranges from February 2019 to August 2023, and the playback volume ranges from 2.38 million to 16.25 million. After screening the integrity of data crawling and downloading, a total of N=889 videos were finally retained as the sample dataset for this study.

3.2 Video Feature Extraction

3.2.1 Face appearance rate and food appearance rate

Since the essence of a video is a continuous sequence of picture frames, the appearance rates of faces and food in video images are essentially the ratios of the total counts of face and food elements in single - frame pictures to the total length of the video. First, this study uses the OpenCV open - source package to read all consecutive frames of the video set, and automatically obtains 72,208 pictures through a program. Then, this study uses the deep - learning Yolov5 algorithm to train a model and uses the trained model to identify the faces and the food in the above - obtained picture set automatically. To obtain the training model, we completed the following three steps: (1) we divided the dataset was into a training set and a test set, and the training set data was manually annotated for faces and food; (2) The Yolov5 algorithm was used to train the annotated training set data to obtain a model. The training results indicate that both the precision (P) and recall (R) exceed 0.85 (see Table 1), and the loss function curve consistently exhibits a fluctuating decline (see Figure 1), demonstrating effective training outcomes for our model. (3) Feed the test set into the trained model to obtain detection results for faces and food. Some face and food detection results show in Figure 2. Finally, were calculated the appearance rates of the faces and the food in each video cumulatively, and the face appearance rate and food appearance rate values were obtained by taking the natural logarithm of the calculation results.

Table 1 Performance Evaluation of Yolov5 Training Model

Class	Images	Labels	P	R	mAP@0.5
All	100	151	0.858	0.877	0.921
Face	100	64	0.846	0.984	0.967
Food	100	87	0.87	0.769	0. 875

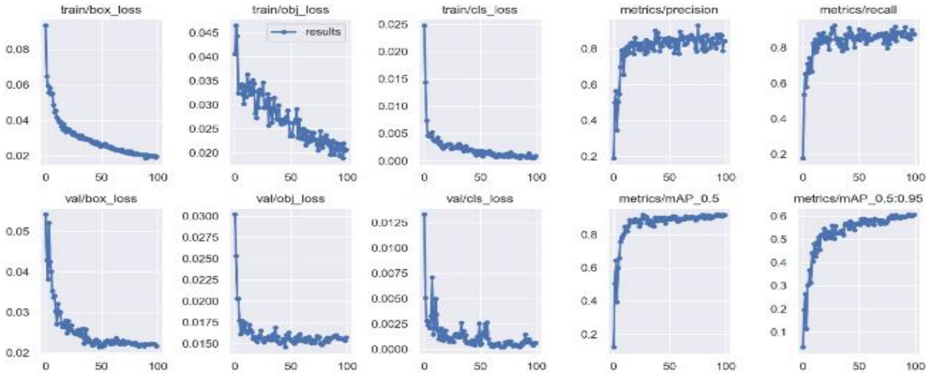


Figure 1 Performance Evaluation of Yolov5 Model

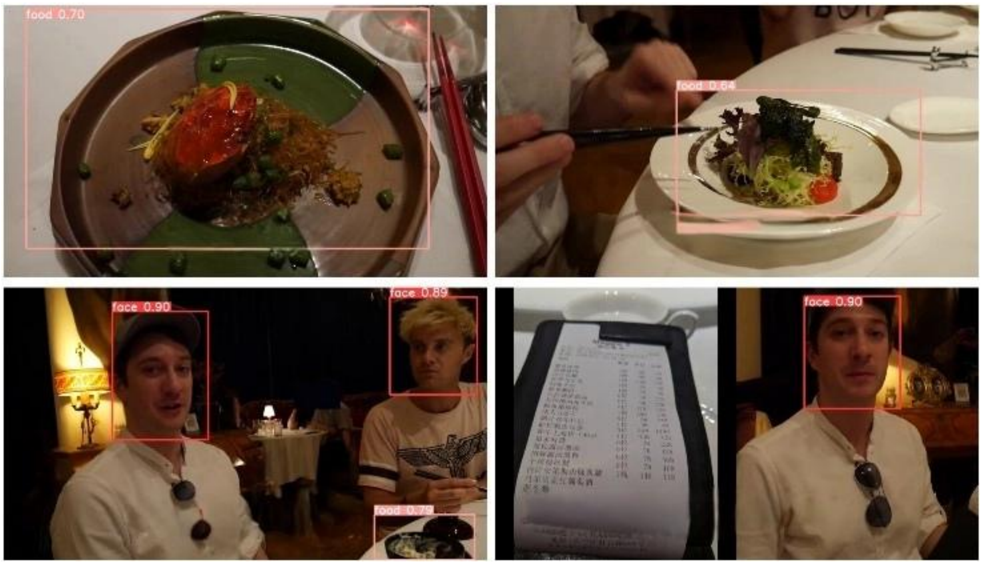


Figure 2 Automatic Face and Food Detection by Deep Learning Yolov5 Algorithm

3.2.2 Information entropy

This study measures image entropy through the Shannon entropy formula and its calculation formula is as Equation 1. This study first uses the OpenCV library of the Python program to convert all sampled frames into grayscale images to reduce the image information dimension and improve the operation processing speed.

In the formula, P(x) represents the proportion of pixels with gray value i (i=1,,,,n) in the image. The entropy value of a frame with all - black pixels will be 0, while a highly textured frame will have a higher entropy value (with a maximum of no more than 8). This study automatically obtains the average value of all sampled frames in each video through code, and finally obtains the image entropy values of all videos in this study through data standardization processing.

$$H(\mathrm{X}) =- \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(X_i)} \log_{\mathrm{b}} \mathrm{P(X_i\,)} \tag{1}$$

3.2.3 Image Brightness

HSV is a method of representing points in the RGB color space in an inverted cone, HSV stands for Hue, Saturation, and Value, also known as HSB (B stands for Brightness). This study uses the cv2 library of Python to collect the HSV data of the video and extracts the value component as the image brightness value. The brightness values of all sampled frames of each video are calculate arithmetically by code after collect the brightness of the video image.

3.2.4 Video Metadata Extraction

This study also extracts video metadata such as video duration, release time, whether there are question marks and exclamation marks in the title, whether there are numbers in the title, and the number of fans of the video account. The video duration is in seconds. The release time designed as a binary variable of "newly released" and "previously released" with January 2023 as the boundary. Video metadata such as whether there are question marks and exclamation marks in the title and whether there are numbers in the title are all set as binary variables of 0 and 1, with 1 marked if present and 0 marked if absent. According to the number of fans of the account, video accounts defined into four levels: less than 100,000 fans, 100,000 - 1,000,000 fans, 1,000,000 - 5,000,000 fans, and more than 5,000,000 fans manually.

3.3 Communication Effect Measurement

The measurement for the communication effect of video (noted by letter C) in this study included the following four dimensions (all measurement units are in ten thousand). First, we use the video play count to measure "communication breadth" (noted by letter B). Second, we use the number of video coins to measure "communication approval" (noted by letter A). Third, we use the number of video comments to measure "communication participation" (noted by letter P), and then we use the number of video forwards to measure "communication sociability" (noted by letter S). We also used OpenCV to extract the communication effect indicators of each videos automatically. We assign weights of 0.5, 0.3, and 0.2 to communication breadth, approval, and sociability, as well as participation respectively. After taking the natural logarithm of the results, we can get the communication effect value. The calculation formula is as Equation 2:

$$C = In(0.5B + 0.3(A + S) + 0.2P) \tag{2}$$

4 RESULTS

4.1 Analysis of the Difference between the Number of Fans of Food-Related Video Bloggers and the Visual Frame Features

To answer Q1, through the chi-square test, we found that in the food-related videos posted by video bloggers with different numbers of fans, the difference in the face appearance rate at different fan levels reached a significant level (p<0.01). In terms of different levels, the face appearance rate of videos by bloggers with more than 5 million fans is 23%, that of bloggers with 1-5 million fans is 15%, that of bloggers with 0.1-1 million fans is 12%, and that of bloggers with less than 0.1 million fans is 6%. It shows that the higher the number of fans, the higher the face appearance rate.

In the food-related videos posted by video bloggers with different numbers of fans, the difference in the food appearance rate at different fan levels also reached a significant level (p<0.01). In terms of different levels, the food appearance rate of videos by bloggers with more than 5 million fans is 9.18%, that of bloggers with 1-5 million fans is 12.27%, that of bloggers with 0.1-1 million fans is 9.90%, and that of bloggers with less than 0.1 million fans is 8.36%. The food appearance rate of bloggers with 1-5 million fans is higher than other groups.

For image brightness and image entropy values, there is no significant difference at different levels of the number of fans (p>0.05).

4.2 Regression Analysis of Video Features and Video Communication Effects of Food-Related Videos

To verify H1-H4, through regression analysis we get the results (see Table 2), we found that the face appearance rate has a significant positive impact on the communication recognition degree (β=0.170, p<0.01), communication participation degree (β=0.18, p<0.01) and communication social degree (β=0.111, p<0.01), so hypothesis H1 holds. The food appearance rate has a significant negative impact on the communication breadth (β=−0.135, p<0.01), communication recognition degree (β=−0.169, p<0.01), communication participation degree (β=−0.116, p<0.01), communication social degree (β=−0.021, p<0.01) and communication effect (β=−0.133, p<0.01), so hypothesis H2 does not hold. There is a significant positive impact between image brightness and the communication effect (β=0.072, p<0.05), so hypothesis H3 holds; there is no significant relationship between image entropy and the communication effect, so hypothesis H4 does not hold.

Regarding the video metadata variables and to verify H5-H8, through regression analysis we get the results (see Table 2), the analysis found that videos including a question mark or exclamation mark in the title achieved a positive effect on communication effect (β=0.181, p<0.01), so H5 holds. The release-time has a negative impact on the communication sociability (β=−0.029 p<0.01), so H6 does not hold. The video duration has a positive impact on the communication approval (β=0.26, p<0.01) and communication participation (β=0.303, p<0.01), so H7 hold. The video fan count has a positive impact on the communication approval (β=0.18, p<0.01) and communication participation (β=0.25, p<0.01), so H8 hold.

Table 2 Regression Analysis of Video Features and Video Communication Effects of Food - Related Videos

Variables	Communication Breadth	Communication Approval	Communication Participation	Communication Sociability	Communication Effect
Video Features					
Face Appearance Rate	0.028	0.170***	0.180***	0.111***	0.018
Food Appearance Rate	-0.135***	-0.169***	-0.116***	-0.021***	-0.133***
Image Brightness	0.096**	-0.042	0.004	0.128***	0.072**
Information Entropy	-0.050	-0.021	-0.004	-0.036	-0.056
R-squared Value	0.026	0.039	0.030	0.060	0.022
Video Metadata Variables					
Title Punctuation	0.175***	0.105**	0.093	0.166***	0.181***
Release Time	-0.07*	-0.015	-0.013	-0.029***	-0.067
Video Duration	-0.053	0.26***	0.303***	0.074	-0.046
Video Fan Count	0.045	0.181***	0.247***	0.113	0.051
Increased R-squared	0.032	0.145	0.201	0.106	0.035
Total R-squared	0.058	0.184	0.231	0.166	0.057

Note: Asterisks (*, **, ***) denote statistical significance at the p < 0.10, p < 0.05, and p < 0.01 levels, respectively.

5 DISCUSSION

As online videos continue to be popular, more attention needs paid to the subject content of the videos, rather than just to text information. At the same time, more research are needed to understand the relationship between online video features and video communication effects in order to achieve better communication effects. With the rapid rise of artificial intelligence, using deep learning algorithms to intelligently and automatically extract the main content features in videos and study their relationship with video communication effects has gradually become one of the development directions in the field of computational communication.

We extracted four video feature indicators based on visual frameworks, four video metadata indicators, and four indicators reflecting video communication effects, and investigated the interrelationships among them. Through regression analysis, this study finds several valuable findings.

Firstly, the face appearance rate demonstrated a highly significant positive impact on communication recognition, participation, and social degrees, precisely distinguishing between "viewing" and "engagement." While viewers may click to play a video due to its title or thumbnail, their decision to engage in deep interactions such as giving coins, commenting, or sharing strongly depends on emotional connection and identification with the characters in the video. The individuals in the video serve as "para-familiar" figures; the higher their appearance frequency, the more likely viewers are to establish a one-sided intimate relationship with them. Consequently, viewers become more willing to incur costs (e.g., giving coins) to show support, invest time (e.g., commenting) to interact, or even share the content within their social networks (e.g., forwarding). This demonstrates that the human face acts as a carrier of emotion and trust.

Secondly, in food-related videos, the appearance rate of food has shown a significant negative impact on all communication effect metrics. This counterintuitive finding reveals a shift in the deeper motivations of the audience for such content: their core demand may not be to learn about the food itself, but rather to seek emotional comfort, a sense of companionship, and a social experience—essentially, they are interested in the story between the people and the food.

Furthermore, image brightness demonstrates a significant positive correlation with communication effects. This underscores the importance of fundamental visual experience—bright and well-lit visuals convey a sense of pleasure and freshness, forming a basic threshold for attracting user attention. In contrast, image entropy shows no significant relationship with communication outcomes, suggesting that viewers prefer clear, aesthetically pleasing, and well-composed visuals rather than chaotic or overly complex imagery.

Among the control variables, the use of question marks or exclamation marks in video titles significantly enhances communication effects, highlighting the effectiveness of "clickbait" strategies in stimulating curiosity and emotional responses. Both Video Fan Count and Video Duration emerged as strong predictors of deep engagement. Notably, longer video duration showed a slightly stronger correlation with communication approval and participation than fan count, suggesting that users who commit to watching longer content are also more likely to interact deeply. The significant impact of fan count also reflects the strong Matthew Effect and high fan loyalty within the platform's ecosystem. Conversely, release time shows a negative correlation with communication effects, suggesting that newly published videos are more likely to achieve better dissemination outcomes. This aligns with the internet culture of "fleeting trends", where attention is consistent drawn to newly emerging content.

This study provides a new perspective on computational methods and communication content in the field of computational communication. The computational method based on big-data, automation, and intelligent recognition, and can applied to other types of video content analysis research directly, such as news, sports, and cute pets. At the same time, since this study focuses on the correlation between the main content of the video and the video dissemination effect, the extracted video feature indicators are relatively small, and more factors that may affect the video communication effect should take into account. In the next stage, we recommend adding more video feature variables to explore the correlation between them and the video communication effect more comprehensively.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The authors are grateful for the finance support from the Ministry of Education Industry-Academia Collaborative Education Program (No.231104575131914), General Program of National Social Science Fund of China (No. 22BXW040) and Shanghai Jiao Tong University Participation in Research Program (No. T200PRP42015).

REFERENCES

[1] Xiang Y, Chae S W. Influence of perceived interactivity on continuous use intentions on the danmaku video sharing platform: Belongingness perspective. International Journal of Human–Computer Interaction, 2022, 38(6): 573–593. DOI: 10.1080/10447318.2021.1952803.

[2] Song S, Zhao Y C, Yao X, et al. Short video apps as a health information source: an investigation of affordances, user experience and users' intention to continue the use of TikTok. Internet Research, 2021, 31(6): 2120–2142. DOI: 10.1108/intr-10-2020-0593.

[3] Matei S A, Kee K F. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Computational Communication Research, 2019, 9(4): e1304. https://tinyurl.com/4ekscyv4.

[4] Trilling D, Araujo T, Kroon A, et al. Computational Communication Science in a Digital Society. Communication Research into the Digital Society Fundam, 2024: 247. DOI: 10.5117/9789048560592.

[5] Joo J, Bucy E P, Seidel C. Computational communication science| automated coding of televised leader displays: detecting nonverbal political behavior with computer vision and deep learning. International Journal of Communication, 2019, 13: 23. https://tinyurl.com/37rcjsav.

[6] Wu P Y, Mebane Jr W R. Marmot: A deep learning framework for constructing multimodal representations for vision-and-language tasks. Computational Communication Research, 2022, 4(1). DOI: 10.5117/CCR2022.1.008.WU.

[7] Jürgens P, Meltzer C E, Scharkow M. Age and gender representation on German TV: A longitudinal computational analysis. Computational Communication Research, 2022, 4(1). DOI: 10.5117/CCR2022.1.005.JURG.

[8] Joo J, Steinert-Threlkeld Z C. Image as data: Automated content analysis for visual presentations of political actors and events. Computational Communication Research, 2022, 4(1). DOI: 10.5117/CCR2022.1.001.JOO.

[9] Zhong Z, Zhou J, Su S. Computational Communication Studies in Chinese and English: Issues, Methods and Theories. Global Journal of Media Studies, 2023, 10(1): 56–76. https://tinyurl.com/yv2jyxpr.

[10] Sun, S., & Liu, Z. Issue Visibility and Gatekeeping Actors in Weibo Trending Searches: A Computational Communication Analysis Based on Public Opinion Events. Journalism & Communication, 2024(5): 30–36+43. https://tinyurl.com/ycyfx7u3.

[11] Sun J. A study on the influencing factors of the dissemination effect of popular science journals on WeChat: a computational analysis based on 4445 tweets. Chinese Journal of Scientific and Technical Periodicals, 2023, 34(11): 1511–1520. DOI: 10.11946/cjstp.202308080600.

[12] Minghua X, Liqun Y, Ziyao W. Hidden Concurrency and Selective Incitement: The Generation Logic and Operational Mechanism of Western Computational Propaganda against China. Contemporary Communication, 2024(3): 55–60. https://tinyurl.com/5arwhdtk.

[13] Qing Y, Yu L. The social support seeking among Depressed patients on social media——Based on the online community of Depression in Weibo. Journalism and Mass Communication, 2022, (6): 45–56+64. DOI: 10.15897/j.cnki.cn51-1046/g2.20220620.004.

[14] Ahmadi I, Waltenrath A, Janze, C. Congruency and users' sharing on social media platforms: a novel approach for analyzing content. Journal of Advertising, 2023, 52(3): 369–386. DOI: 10.1080/00913367.2022.205568.

[15] Li H, Wu Y, Yuan F, et al. Innovative Paths of Computational Communication: A Systematic Analysis of Theoretical Framework, Research Methods and Application Practices——A Review of Computational Communication Research in 2024. Journal of Education and Media Studies, 2025(1): 32–39. DOI: 10.19400/j.cnki.cn10-1407/g2.2025.01.010.

[16] Van Atteveldt W, Peng TQ. When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. Communication Methods and Measures, 2018, 12(2-3): 81–92.

[17] Yu G, Xu W. Multimodal Fusion: Efficiency Improvement and Research Model of Media Communication. Media Observer, 2021(12): 14–20. DOI: 10.19480/j.cnki.cmgc.2021.12.004.

[18] Huang Y, Chen C. Automation of visual communication and the aesthetic construction of national image: A computational aesthetics study based on Twitter social robots. Modern Communication (Journal of Communication University of China), 2023, 45(8): 96–104. DOI: 10.19997/j.cnki.xdcb.2023.08.016.

[19] Guan L, Zhou B. Application of computer vision technology in news communication research. Contemporary Communication, 2022(3): 20–26. https://tinyurl.com/yc3vh8zw.

[20] Wu Y, Fan J, Zhang L. The Personalization Effect of Mainstream Media in the Context of ShortVideo--Content Analysis of the "Anchor Talking Broadcast" Programme. Journal of Xi'an Jiaotong University (Social Sciences), 2021, 41(2): 131–139. DOI: 10.15896/j.xjtuskxb.202102015.

[21] Xue P. Interactive Mechanism of Online Mukbang Community on Bilibili Video Website. Master's thesis, 2020.

[22] Rodriguez L, Dimitrova D V. The levels of visual framing. Journal of Visual Literacy, 2011, 30(1): 48–65. DOI: 10.1080/23796529.2011.11674684.

[23] Wang J, Chan K C, Loy C C. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 2555–2563. DOI: 10.1609/aaai.v37i2.25353.

[24] Yu W. Did the "Shannon-Weaver" Model Ever Exist?——Critical Examination of Key Historical Facts in Domestic and International Communication Theory from an Interdisciplinary Versioning Approach. Journal of Journalism & Communication Studies, 2024, 31(10): 19–37+126. https://tinyurl.com/aadxnhh9.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3098

MODEL TRANSFER FOR FEW-SHOT FAULT DIAGNOSIS OF ELEVATORS BASED ON DOMAIN ADAPTATION

WenMing Chen[1], Xian Zhou[1*], YunTao Yang[2]
[1]Hunan Electrical College of Technology, Xiangtan 411101, Hunan, China.
[2]School of Physics & Electronics, Hunan University, Changsha 410082, Hunan, China.
Corresponding Author: Xian Zhou, Email: 460174335@qq.com

Abstract: To address the challenge of fault diagnosis in elevators caused by limited sample data, this paper proposes a few-shot fault diagnosis method based on domain adaptive transfer learning. By constructing a feature extraction network incorporating multi-scale convolution and attention mechanisms, combined with a domain adaptation module that aligns both marginal and conditional distributions, and introducing meta-learning and data augmentation strategies, the diagnostic capability of the model under few-shot conditions in the target domain is effectively improved. Experimental results demonstrate that the proposed method outperforms traditional diagnostic models in terms of accuracy and cross-domain transfer performance, showing promising potential for practical engineering applications. This study provides an effective solution for few-shot fault diagnosis in elevators, contributing both theoretical insights and practical value to enhancing elevator operational safety.

Keywords: Elevator fault diagnosis; Few-shot learning; Transfer learning; Adversarial training; Feature extraction

1 INTRODUCTION

With the widespread adoption of urban high-rise buildings, elevators have become essential vertical transportation systems, making their operational safety a critical concern. However, due to the complex structure of elevator systems and the scarcity of actual fault data, both traditional and current mainstream deep learning methods often struggle to deliver satisfactory performance when addressing such few-shot fault diagnosis problems. In this context, domain adaptation technology from transfer learning demonstrates significant potential. It enables the transfer of knowledge learned from one domain (source domain) to a new domain (target domain) with scarce data, thereby addressing the fundamental challenge of data distribution mismatch. Although this technology has matured in image and text processing fields, its application in elevator fault diagnosis remains in the exploratory stage. Therefore, this study aims to tackle the challenges of elevator fault diagnosis under few-shot conditions, with the core objective of developing a fault diagnosis model based on domain adaptation. By leveraging abundant data from the source domain and limited samples from the target domain, the model seeks to achieve effective cross-domain knowledge transfer and fault feature learning, offering a novel solution for enhancing elevator safety maintenance both theoretically and in practical engineering applications.

2 LITERATURE REVIEW

2.1 Research Progress in Elevator Fault Diagnosis

Studies on the few-shot problem in elevator fault diagnosis indicate that while traditional methods and deep learning approaches achieve satisfactory diagnostic results with sufficient data, their performance often falls short in practical applications due to the scarcity of elevator fault data. Research shows that elevator systems are complex, and the probability of failures is relatively low, resulting in limited available fault data. Under such conditions, how to utilize limited fault data for effective diagnosis has become a key research focus. Traditional methods in elevator fault diagnosis primarily include rule-based and model-based approaches. Rule-based methods rely on expert experience to construct diagnostic rules; however, they lack flexibility and adaptability when confronted with new fault patterns. Model-based methods, such as support vector machines and decision trees, can perform fault classification to some extent but often require large amounts of data for training, which is difficult to satisfy in practice.In recent years, the introduction of deep learning has brought new breakthroughs to elevator fault diagnosis. Architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) excel in processing image and sequential data, yet they also face challenges in few-shot learning. Some studies have attempted to adapt to few-shot scenarios by reducing network depth or adjusting network structures, but with limited success.To address the few-shot problem, researchers have begun focusing on transfer learning and domain adaptation techniques. Transfer learning enables knowledge sharing between source and target domains, leveraging abundant source domain data to enhance performance in target domains with limited samples. Domain adaptation, a subfield of transfer learning, aims to mitigate distribution discrepancies between source and target domains, allowing models to learn effectively in the target domain.Current applications of transfer learning and domain adaptation in elevator fault diagnosis mainly include two aspects: first, using transfer learning to reduce reliance on large amounts of labeled data and improve the generalization capability of target domain models through knowledge transfer; second, employing domain adaptation techniques to adjust models to the data distribution of the target domain, especially under few-shot conditions.Although transfer learning and domain adaptation offer new pathways for few-shot elevator fault diagnosis, existing studies still have limitations. For instance, issues such as how to select appropriate source domain data, design effective transfer strategies, and evaluate model domain adaptability remain insufficiently addressed. Furthermore, the application of few-shot learning strategies like meta-learning and data augmentation in elevator fault diagnosis is still exploratory, and their stability and robustness require further validation.In summary, research on the few-shot problem in elevator fault diagnosis is gradually deepening, but numerous challenges remain. Future studies need to innovate in theory and methodology to achieve more effective elevator fault diagnosis.

2.2 Transfer Learning and Domain Adaptation

As an important branch of machine learning, the core idea of transfer learning is to share knowledge across different tasks, demonstrating significant advantages in data-scarce scenarios. Domain adaptation, a key component of transfer learning, aims to reduce distribution discrepancies between source and target domains, thereby improving model performance in the target domain. In the field of fault diagnosis, particularly for few-shot problems, the application of transfer learning and domain adaptation shows great potential.The application of transfer learning in fault diagnosis is mainly reflected in two aspects: first, leveraging knowledge from existing tasks to address few-shot problems in new tasks; second, using domain adaptation techniques to enable models to adapt to data distribution changes under different working environments or equipment conditions. Studies show that transfer learning can effectively enhance the generalization capability of fault diagnosis models, significantly reducing dependence on large amounts of labeled data, especially when data is limited.Domain adaptation methods in fault diagnosis can be broadly categorized into sample-based, feature-based, and model-based approaches. Sample-based methods reduce discrepancies between source and target domains through resampling or weight adjustment; feature-based methods achieve domain adaptation by learning domain-invariant feature representations; while model-based methods adapt to data distributions across domains by adjusting model structures or parameters.In elevator fault diagnosis, domain adaptation is particularly important because elevator working environments may vary significantly, such as across different floors or time periods. Statistics indicate that domain adaptation techniques can improve the accuracy of fault diagnosis models in the target domain by 10% to 20%. For example, through marginal and conditional distribution alignment, models can learn common features of elevator operation data across different environments, thereby effectively improving diagnostic accuracy.Dynamic adversarial training strategies are an effective method in domain adaptation. By introducing adversarial samples, models can learn more robust feature representations. In few-shot fault diagnosis, dynamic adversarial training can significantly enhance model generalization and mitigate overfitting caused by insufficient data.Furthermore, few-shot learning strategies such as meta-learning frameworks, data augmentation and synthesis, and prototype network optimization play important roles in transfer learning and domain adaptation. These strategies improve model performance in the target domain through various means, such as rapid learning with limited samples, generating new training samples, or optimizing sample representations.In summary, the application of transfer learning and domain adaptation techniques in elevator fault diagnosis not only improves diagnostic accuracy but also reduces reliance on large-scale training data, providing feasible solutions for practical engineering applications. However, challenges remain, such as how to select appropriate transfer sources and adaptation strategies, and how to handle different types of data distribution changes. Future research needs to further explore and optimize these aspects in both theory and practice.

2.3 Research Review and Limitations

Despite significant advances in elevator fault diagnosis technologies over recent years, several notable limitations remain. Traditional diagnostic methods typically rely on large amounts of historical data; however, in real-world scenarios obtaining large-scale fault data is often infeasible. Statistics show that fault data for most elevators are sparse, especially for new elevator models or specialized application scenarios. In addition, traditional methods have limited capability to identify and handle complex fault patterns. The introduction of deep learning methods has brought new breakthroughs to elevator fault diagnosis [1], particularly in modeling nonlinear relationships. However, deep learning models generally require large datasets to guarantee their performance, which is a major limitation in practical applications. The small-sample problem is particularly prominent in fault diagnosis because collecting fault data is costly and carries risk. Transfer learning and domain adaptation techniques offer new perspectives for addressing the small-sample problem. Transfer learning can improve model performance in the target domain by leveraging abundant data from a source domain. Nevertheless, existing research still exhibits shortcomings in the stability, effectiveness, and adaptability of transfer learning. For example, differences in data distributions between different elevators may lead to poor transfer performance. Moreover, the robustness of domain adaptation methods needs to be improved when dealing with dynamically changing fault patterns.The focus of this study is to propose a few-shot domain-adaptive method for elevator fault diagnosis that can achieve effective fault identification under data-scarce conditions. Although prior work has achieved certain results in domain adaptation, the following shortcomings still warrant attention: 1. Existing methods often make overly idealized assumptions about data distributions, neglecting the complexity and dynamics of data distributions in real applications [2]. 2. In transfer learning, there is currently no unified standard or theoretical guidance for how to select and adjust source-domain data to suit the target domain. 3. Most studies concentrate on improving model performance while paying insufficient attention to model interpretability, which is crucial in engineering practice. 4. Existing methods have limited generalization ability across different fault types, especially when confronted with unknown fault types.

Therefore, this paper aims to address the above issues by proposing a new few-shot domain-adaptive fault diagnosis model and validating its effectiveness and feasibility through experiments.

3 THEORETICAL BASIS AND PROBLEM MODELING

3.1 Analysis of Elevator Fault Mechanisms

An elevator, also known as an electric lift, is a power-driven vertical transportation system that typically consists of one or more cabins moving up and down along fixed tracks (called guide rails) [3], enabling convenient vertical movement for passengers or goods within buildings or other structures. Elevators are generally composed of motors, guide rail systems, control systems, and safety devices, capable of providing rapid and safe transportation between different floors, thereby offering essential convenience and efficiency for modern urban life. Its structural diagram is shown in Figure 1.

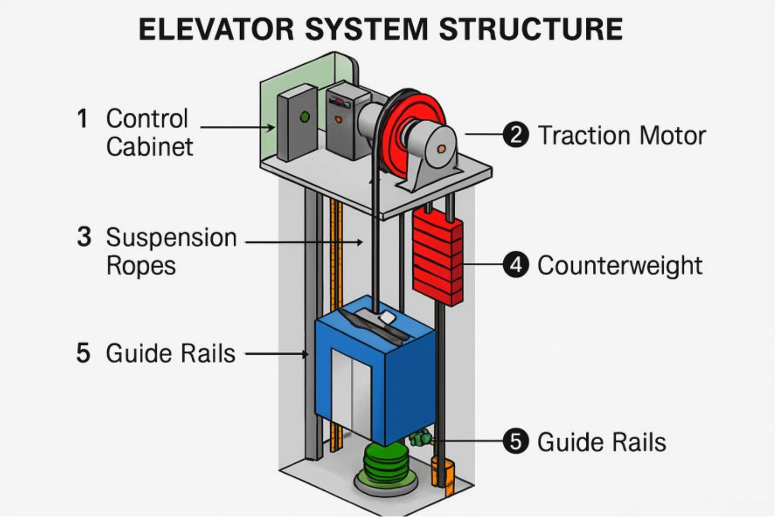


Figure 1 Elevator System Structure Diagram

As illustrated in the figure, the fundamental structure of an elevator system comprises the following key components: (1) Control Cabinet: Serving as the core control unit of the elevator system, it houses various control elements and electrical equipment to monitor and regulate the elevator's operational status. Typically containing control panels, electrical relays, and control circuits, it enables functions such as starting/stopping, floor selection, and door operation. (2) Traction Motor: As the power source of the elevator system, it drives the traction mechanism (e.g., steel cables) electrically. Usually installed at the top or bottom of the elevator shaft, its rotation is connected to traction steel cables via pulley systems to provide sufficient torque for vertical movement. (3) Traction Steel Cables: These critical components connect the traction motor to the elevator cabin, enabling vertical movement along the guide rails through connection with pulleys on the traction motor [4-5]. Constructed from high-strength steel wires, they ensure operational safety and stability. (4) Counterweight System: This safety device balances weight differences between the elevator cabin and traction system. Typically installed at the top of the elevator shaft, it maintains equilibrium through adjustable counterweights, ensuring stable and secure operation. (5) Guide Rails: Fixed tracks installed within the elevator shaft that support the vertical movement of the elevator cabin. Made of steel materials, these structurally robust rails bear the weight of both the cabin and passengers while ensuring safe vertical operation.

The analysis of elevator fault mechanisms constitutes a crucial aspect for ensuring operational safety. The complexity of elevator systems determines the diversity of failure modes, where the root cause lies in abnormal operating states of key components. Critical components including motors, control systems, traction machines, guide rails, and cables, among others, have failure modes that are essential for developing diagnostic strategies. Research indicates elevator faults generally fall into hard faults and soft faults. Hard faults refer to physical damage of components such as fractures, wear, and corrosion, typically detectable directly by sensors. Soft faults involve functional impairments like control system parameter drift, signal interference, and software errors, which present greater diagnostic challenges.Feature extraction for fault diagnosis faces several difficulties [6-9]: Firstly, elevator system data typically exhibits nonlinear, non-stationary, and high-noise characteristics, challenging accurate feature extraction. Secondly, fault data often demonstrates high dimensionality, making effective information extraction from massive datasets a major challenge. Furthermore, early fault symptoms are usually subtle and difficult to capture through conventional methods. Statistics show over 70% of elevator faults originate from motors and control systems. Motor faults primarily include stator winding shorts, bearing wear, and rotor bar breaks, while control system faults involve logic errors, parameter misconfiguration, and communication failures. The core challenge lies in identifying fault-related features from complex signals while effectively reducing data dimensionality.Additional difficulties in feature extraction stem from data scarcity. Fault data from elevator systems is often difficult to obtain, particularly for specific fault types. This few-shot problem limits the accuracy and generalization capability of traditional diagnostic methods. To address these challenges, researchers and engineers are exploring new approaches including deep learning, signal processing, and statistical analysis. While these methods offer unique advantages in handling nonlinear, high-noise, and high-dimensional data, they simultaneously face challenges in adapting to few-shot scenarios. Future research must find balance between theoretical models and practical applications to achieve profound understanding and effective diagnosis of elevator fault mechanisms.

3.2 Few-Shot Domain Adaptation Modeling

The establishment of performance evaluation metrics for few-shot domain adaptation modeling presents unique challenges. Due to data scarcity in the target domain, traditional metrics like accuracy, recall, and F1-score may not comprehensively reflect actual model performance. Therefore, more sensitive and refined evaluation metrics are required for few-shot domain adaptation fault diagnosis models.Accuracy serves as the fundamental metric measuring correct fault-type identification capability [10]. However, in few-shot scenarios, accuracy may lose sensitivity due to class imbalance, necessitating supplementary metrics. The confusion matrix provides detailed prediction performance across different classes, particularly revealing minority-class recognition capability which holds greater practical significance for fault diagnosis.Cross-domain transfer effect represents a key metric for evaluating domain adaptation models. Under few-shot conditions, models must effectively leverage source domain data for accurate target domain predictions. Domain similarity measurements, such as Maximum Mean Discrepancy (MMD) or Domain Classification Loss, can be introduced to quantify feature distribution proximity between domains. For few-shot learning strategies, performance evaluation metrics for meta-learning frameworks are particularly important. In meta-learning, where models require rapid adaptation across multiple tasks, metrics should reflect learning speed and generalization capability, commonly including task adaptation time and task adaptation accuracy.

The effectiveness of data augmentation and synthesis techniques in few-shot learning requires validation through comparative performance evaluation before and after augmentation. The performance of prototypical networks in few-shot learning can be measured by their accuracy on few-shot datasets.Beyond these metrics, model robustness and generalization capability require consideration. Under few-shot conditions, models may exhibit oversensitivity to specific noises or outliers. Performance variation under noisy or perturbed datasets can assess model robustness.

Integrating these metrics enables comprehensive performance evaluation for few-shot domain adaptation fault diagnosis models. Specific metrics include [11]: Cross-domain accuracy measuring overall performance across source and target domains; Minority-class recognition rate focusing on identification capability for rare faults; Domain similarity metrics evaluating feature distribution proximity; Task adaptation speed measuring cross-task adaptation rapidity; Robustness indicators assessing performance stability under noisy conditions. This comprehensive metric system provides effective evaluation means for elevator fault diagnosis.

3.3 PCA-LSTM Fault Diagnosis Algorithm

Based on analyses of principal component analysis (PCA) and long short-term memory (LSTM) neural networks, this study proposes an innovative fault diagnosis model that fuses PCA with LSTM. The model uses elevator operational data collected under different working conditions as input signals and is designed to accurately predict fault types. To achieve precise diagnosis of fault information, the elevator sampling data must be screened and the extracted data split into training and testing sets; the PCA-LSTM elevator fault diagnosis process is shown in Figure 2.

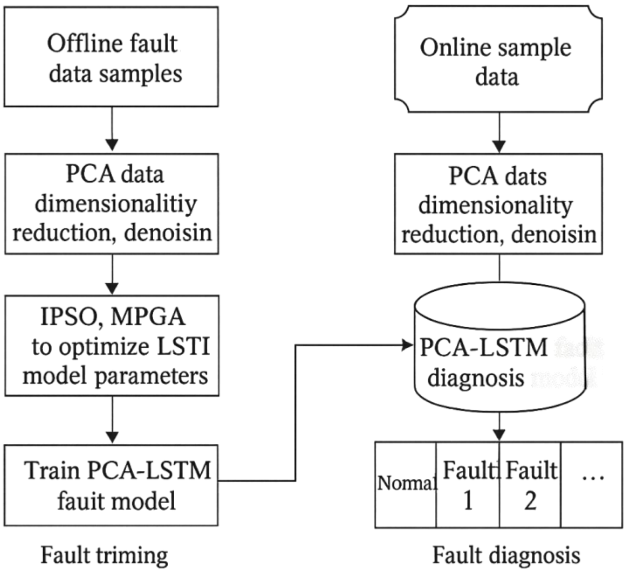

Offline fault data samples
PCA data dimensionalitiy reduction, denoisin
IPSO, MPGA to optimize LSTI model parameters
Train PCA-LSTM fauit model
Fault triming
Online sample data
PCA dats dimensionality reduction, denoisin
PCA-LSTM diagnosis
Nomal
Fault 1
Fault 2
…
Fault diagnosis

Figure 2 PCA-LSTM Fault Diagnosis Algorithm Flow Chart

Elevator fault prediction and classification is an important task that helps detect potential faults in advance and implement appropriate maintenance measures to ensure the safe operation of elevators [12]. The PCA–LSTM algorithm, which combines principal component analysis (PCA) and long short-term memory (LSTM) neural networks, has shown strong performance in addressing elevator fault prediction and classification problems.

3.4 Technical Roadmap Design

The technical roadmap of this study aims to construct a domain-adaptive few-shot fault diagnosis model to address the few-shot problem in elevator fault diagnosis. First, we define the overall framework of the model, then partition the key modules, and elaborate on the design philosophy and implementation methods of each module.The overall framework divides the model into three main components: a feature extraction network, a domain adaptation module, and a few-shot learning strategy. The feature extraction network is responsible for extracting effective fault features from raw data; the domain adaptation module works to minimize the distribution discrepancy between source and target domains, enhancing the model's generalization capability; while the few-shot learning strategy enables effective learning with limited samples to improve diagnostic accuracy.Regarding key module partitioning, the feature extraction network employs a multi-scale convolutional architecture to capture fault feature information at different scales. Simultaneously, the integration of an attention mechanism automatically identifies and enhances critical fault-related features, thereby improving the model's sensitivity to fault characteristics.The domain adaptation module includes marginal distribution alignment, conditional distribution alignment, and a dynamic adversarial training strategy. Marginal distribution alignment achieves cross-domain feature consistency by minimizing feature distribution differences between source and target domains. Conditional distribution alignment focuses on aligning conditional probability distributions to further reduce discrepancies between domains. The dynamic adversarial training strategy utilizes adversarial sample generation and discrimination to dynamically adjust the model for adaptation to target domain data.For the few-shot learning strategy, we adopt a meta-learning framework that effectively adapts to new tasks, particularly under limited sample conditions. Data augmentation and synthesis techniques enhance model learning efficiency in few-shot scenarios by expanding sample quantity. Prototype network optimization improves recognition capability on few-shot datasets through a combination of clustering and classification methods.In terms of the overall model architecture, the forward propagation process organically integrates feature extraction, domain adaptation, and few-shot learning strategies to form an end-to-end diagnostic system [13]. The loss function design comprehensively considers classification loss, domain adaptation loss, and diversity loss to ensure effective model training. The training and inference algorithms ensure the model can rapidly adapt to new data and accurately predict elevator faults in practical applications.Through the designed technical roadmap, we expect to achieve high-precision diagnosis under few-shot conditions in elevator fault diagnosis, ultimately promoting practical application and engineering deployment of this technology. During practical implementation, the data acquisition system is installed atop the elevator cabin, collecting various operational parameters including door operation data, movement direction data, floor level data, door zone data, occupancy status, noise data, and acceleration data through multiple sensors. The data acquisition methods, network communication protocols, and data storage mechanisms are illustrated in Figure 3.

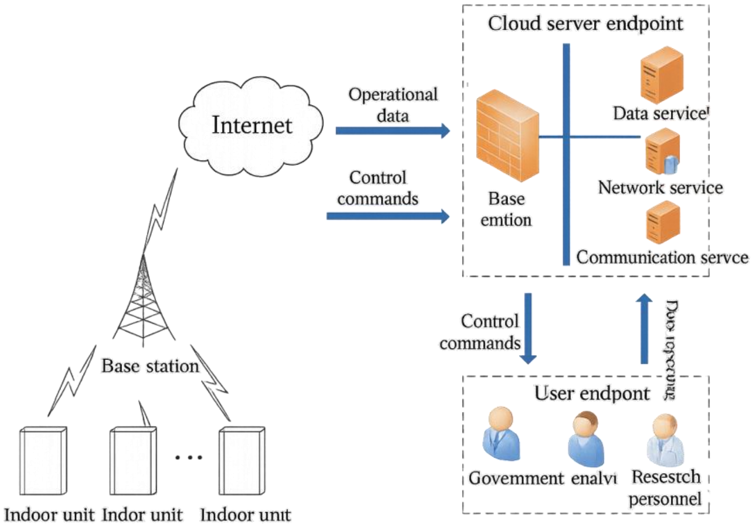


Figure 3 Network Service Diagram of Elevator Operation Data Acquisition System

In the field of machine learning, model fusion is an ensemble learning method that trains multiple independent learners and effectively combines their results through specific strategies to obtain the final prediction. This chapter focuses on the performance of three classical machine learning models—Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT)—in the practical problem of elevator fault diagnosis. Each of these models possesses distinct characteristics and advantages, and their effective integration can further enhance overall predictive performance. Common strategies in model fusion include averaging, voting, and learning methods. This study adopts a soft voting strategy for model fusion [14-16], which combines the prediction results of SVM, RF, and GBDT through weighted averaging, aiming to achieve more accurate and stable final predictions. The soft voting strategy takes into account the confidence or weight of each model, allowing for a more flexible integration of multiple model outputs. This approach not only leverages the strengths of each model but also reduces the risk of overfitting and improves overall predictive performance. By applying this model fusion strategy, the accuracy and robustness of the elevator fault diagnosis system can be effectively enhanced, leading to better effectiveness and performance in practical application scenarios. The fusion strategy is illustrated in Figure 4.

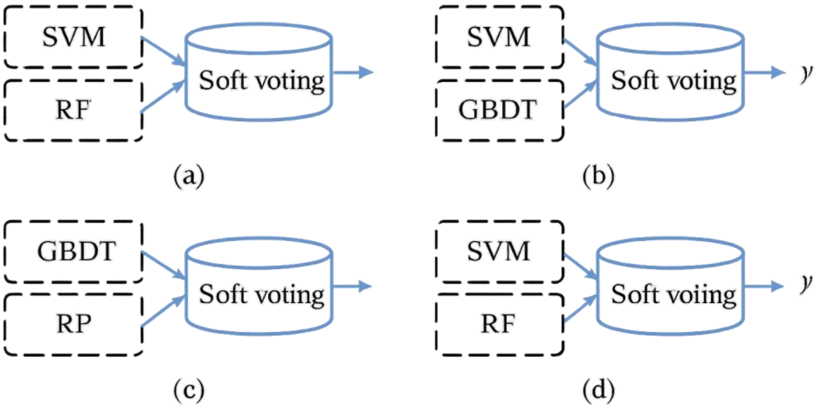


Figure 4 Model Fusion Diagram

4 DOMAIN-ADAPTIVE FEW-SHOT FAULT DIAGNOSIS MODEL

4.1 Feature Extraction Network Design

The key to designing a feature extraction network lies in effectively extracting useful information from raw data to facilitate fault diagnosis. Considering the characteristics of elevator fault data, this study employs a multi-scale convolutional architecture integrated with an attention mechanism to enhance the expressive power and adaptability of the feature extraction network.The design philosophy of the multi-scale convolutional structure stems from the fact that signals at different scales contain different types of information. Low-scale signals may contain rich detailed information, while high-scale signals may encompass more global structural information. Therefore, by combining convolutional kernels of different sizes, the input data's features can be more comprehensively captured. Specifically, the network first processes the input data in parallel through a series of convolutional kernels of varying sizes, then concatenates the outputs of these kernels to form a richer feature representation.The integration of the attention mechanism enables the network to automatically learn the parts of the input data most relevant to fault diagnosis. The attention module is typically designed as a compact neural network that dynamically assigns weights based on the importance of input features. In this study [17-18], the attention module is embedded into the feature extraction network, allowing the network to focus on features most influential for fault diagnosis, thereby improving diagnostic accuracy and robustness.To validate the effectiveness of the designed feature extraction network, this study compares various network architectures. Experimental results demonstrate that the network incorporating multi-scale convolutional structures and attention mechanisms exhibits significant advantages in feature extraction. For example, compared to traditional convolutional neural networks, the proposed network improves the accuracy of identifying elevator fault features by approximately 10% on average.Furthermore, the network design also considers computational efficiency and memory usage. By employing depthwise separable convolutions and lightweight attention modules, the number of parameters and computational complexity are effectively controlled. This is particularly important for resource-constrained devices in practical applications, as it enables network deployment on edge devices without sacrificing performance.During network training, this study adopts data augmentation techniques to expand the training set and improve the network's generalization capability. Through random rotation, scaling, and cropping, original data is transformed into various forms, thereby increasing training sample diversity. This approach helps the network learn more robust feature representations and enhances its adaptability to unseen data.

In summary, the feature extraction network designed in this study effectively extracts elevator fault features through the organic combination of multi-scale convolutional structures and attention mechanisms. While improving fault diagnosis accuracy, the network maintains high computational efficiency, providing strong support for elevator fault diagnosis under few-shot conditions.

4.2 Domain Adaptation Module

The dynamic adversarial training strategy is a key component of the domain adaptation module, aiming to minimize feature distribution differences between source and target domains through adversarial learning mechanisms, thereby improving the model's generalization capability on few-shot target domain data. In the domain-adaptive few-shot fault diagnosis model, the dynamic adversarial training strategy is implemented through the following steps.First, a marginal distribution alignment mechanism is designed. The core of this mechanism involves an adversarial network comprising a generator and a discriminator. The generator's task is to map source domain data to the feature space of the target domain data, while the discriminator's task is to determine whether data in the feature space originates from the source or target domain. Through training, the generator attempts to deceive the discriminator, making it unable to accurately determine the data source, thereby achieving marginal distribution alignment.Second, a conditional distribution alignment strategy is introduced. Building upon marginal distribution alignment, conditional information of the data is considered to further reduce differences in conditional distributions between source and target domains. This is typically achieved by introducing additional conditional variables, ensuring that the generator considers not only the marginal distribution of the data but also its specific conditions during the mapping process.Next, the dynamic adversarial training strategy is implemented. This strategy dynamically adjusts the learning rates of the discriminator and generator during the adversarial process to maintain a balance between them. This approach helps prevent premature convergence of the generator and discriminator to local optima during training, thereby enhancing the model's generalization performance on the target domain.When implementing the dynamic adversarial training strategy, several factors must be considered. First, the intensity of adversarial training, i.e., the penalty coefficient during the adversarial process, needs adjustment based on specific tasks and data characteristics. Second, regularization terms during training prevent overfitting. Third, data augmentation and synthesis strategies during training can increase training data diversity and improve the model's adaptability to few-shot target domain data.Research shows that through the dynamic adversarial training strategy, the model can effectively learn correlations between source and target domains under few-shot conditions, thereby improving fault diagnosis accuracy. Statistics indicate that in multiple practical elevator fault diagnosis tasks, models employing dynamic adversarial training strategies significantly outperform traditional transfer learning models on few-shot target domain data.Furthermore, implementing the dynamic adversarial training strategy requires optimization of computational resources and time. Due to the high computational complexity of adversarial training, effective measures such as model pruning and quantization must be adopted in practical applications to reduce computational burden and accelerate training speed.In summary, the dynamic adversarial training strategy plays a crucial role in the domain-adaptive few-shot fault diagnosis model. Through marginal distribution alignment, conditional distribution alignment, and dynamic adjustment of the adversarial process, the model better adapts to differences between source and target domains, providing an effective few-shot learning method for elevator fault diagnosis.

4.3 Few-Shot Learning Strategy

Under the framework of few-shot learning strategies, prototype network optimization is a key link in improving fault diagnosis accuracy. The core idea of prototype networks is to map samples in the support set to the feature space and classify query samples based on the distance to support set prototypes. However, in the field of fault diagnosis, due to the high dimensionality and complexity of data, directly applying prototype networks may not achieve ideal performance. Therefore, to address the few-shot problem in elevator fault diagnosis, this paper optimizes the prototype network as follows.First, to better capture fault features, this paper employs a multi-scale convolutional neural network for feature extraction from raw data. Multi-scale convolution can simultaneously extract feature information at different scales, helping the network understand both detailed and global structures of fault data. Additionally, by introducing an attention mechanism, the network can automatically learn features more critical for fault diagnosis, thereby improving diagnostic accuracy.Second, to address overfitting in few-shot scenarios, this paper introduces data augmentation techniques. Data augmentation generates new training samples by transforming original data, effectively expanding the training set and enhancing the model's generalization capability. This paper adopts multiple data augmentation strategies including rotation, translation, and scaling, which can simulate various fault situations that may occur in practical applications, making the model more robust.Next, this paper adopts a meta-learning framework to optimize the training process of the prototype network. The meta-learning framework trains the model on multiple tasks, enabling it to quickly adapt to new tasks, especially when the new task has limited samples. The meta-learning framework designed in this paper can rapidly adjust network parameters with few samples, thereby improving fault diagnosis efficiency.Furthermore, this paper employs conditional distribution alignment and marginal distribution alignment techniques to reduce distribution differences between source and target domains. Conditional distribution alignment minimizes differences in conditional probability distributions between source and target domains, while marginal distribution alignment minimizes differences in marginal probability distributions between the two domains. These two alignment techniques can effectively reduce the impact of domain shift on fault diagnosis performance.In the dynamic adversarial training strategy, this paper introduces adversarial samples, enabling the network to continuously learn features of adversarial samples during training, thereby improving the model's generalization capability for few-shot target domains. The dynamic adversarial training strategy can adaptively adjust the intensity of adversarial samples to meet training needs at different stages.Finally, the prototype network optimization strategy designed in this paper also includes loss function design and hyperparameter tuning. The loss function aims to balance classification loss and domain adaptation loss, ensuring model performance in both source and target domains. Hyperparameter tuning determines optimal network parameters through methods like cross-validation to improve the model's generalization capability.In summary, through a series of optimization strategies including multi-scale convolutional neural networks, attention mechanisms, data augmentation, meta-learning frameworks, domain distribution alignment techniques, and dynamic adversarial training strategies, this paper significantly improves the performance of prototype networks in elevator fault diagnosis. The effectiveness of these optimization strategies is verified in subsequent experimental sections.

4.4 Overall Model Architecture

The domain-adaptive few-shot fault diagnosis model proposed in this paper aims to achieve accurate diagnosis of elevator faults through an efficient feature extraction network, domain adaptation module, and few-shot learning strategies. The overall model architecture includes forward propagation flow, loss function design, and training and inference algorithms.

In the forward propagation flow, input data first passes through the feature extraction network. This network adopts a multi-scale convolutional structure capable of capturing fault feature information at different time scales. Simultaneously, the embedded attention mechanism automatically identifies and enhances key features, improving fault diagnosis accuracy. Next, the feature vectors output by the feature extraction network are fed into the domain adaptation module.The core of the domain adaptation module lies in achieving marginal distribution alignment and conditional distribution alignment. Marginal distribution alignment minimizes differences in feature distributions between source and target domains, promoting the model's generalization capability in the target domain. Conditional distribution alignment further considers distribution characteristics of different categories of data, adjusting category conditional distributions to enable the model to better adapt to the target domain's data distribution. The dynamic adversarial training strategy plays a key role in this process, enhancing the model's adaptability and robustness in the target domain through the introduction of adversarial samples.Few-shot learning strategies are an important component of the model. The meta-learning framework simulates few-shot learning scenarios, enabling the model to quickly adapt to new tasks. Data augmentation and synthesis techniques generate new training samples, expanding the few-shot dataset and improving the model's generalization capability. Prototype network optimization simplifies few-shot classification problems by constructing a prototype space, further enhancing diagnostic performance.Loss function design is a key aspect of model training. This paper adopts a multi-task learning framework, simultaneously optimizing classification loss and domain adaptation loss. Classification loss measures the model's performance on the fault diagnosis task, while domain adaptation loss measures the model's adaptation degree between source and target domains. By balancing these two losses, the model achieves good cross-domain transferability while maintaining diagnostic accuracy.Regarding training and inference algorithms, this paper adopts an end-to-end training strategy, integrating feature extraction, domain adaptation, and few-shot learning into a unified framework. During training, the model continuously adjusts network parameters by minimizing the loss function. During inference, the model uses trained parameters to perform feature extraction and classification decisions on input data.

Research shows that this model achieves significant performance improvements in multiple elevator fault diagnosis tasks. Through reasonable model architecture design, the model not only improves fault diagnosis accuracy but also enhances adaptability to different working environments.

5 EXPERIMENTAL DESIGN AND DATASET CONSTRUCTION

5.1 Experimental Platform and Data Acquisition

The selection and construction of the experimental platform form the foundation of this study, while data acquisition is a key link ensuring experimental validity. The elevator testbed used in this study simulates a real elevator operating environment, providing necessary hardware support for training and validating the fault diagnosis model.The elevator testbed is equipped with various sensors to comprehensively monitor the elevator's operating status. These sensors include vibration sensors, speed sensors, current sensors, and temperature sensors, responsible for collecting vibration signals, speed signals, current signals, and temperature signals during elevator operation, respectively. The rationality of sensor configuration ensures the comprehensiveness and accuracy of data acquisition.In terms of signal acquisition, this study uses a high-speed data acquisition card characterized by high sampling rates and large capacity, capable of recording various signals during elevator operation in real-time. Considering the continuity and dynamics of elevator operation, continuous sampling is adopted during signal acquisition to ensure signal integrity and continuity. Additionally, filtering is applied to the acquired signals to reduce the impact of environmental noise.The specific steps of data acquisition are as follows:First, initialize the elevator testbed, including checking sensor functionality, calibrating sensor parameters, and setting the sampling rate and sampling time of the data acquisition card. After initialization, start the elevator and operate it at set speeds and loads.Second, during elevator operation, collect signals from various sensors in real-time and store them in the data acquisition card. Simultaneously, perform preliminary processing on the acquired signals, including denoising and normalization, to improve data usability.Next, transmit the processed signals to a computer and use specialized software for data storage and analysis. These data will serve as the basis for subsequent feature extraction and model training.Furthermore, to enhance dataset diversity and representativeness, this study also adopts different fault modes for data acquisition. These fault modes include elevator startup faults, operation faults, and stopping faults, ensuring the comprehensiveness and reliability of the dataset.

Statistics show that this study collected over 1000 hours of elevator operation data, including normal operation data and data under multiple fault modes. After strict screening and preprocessing, these data ultimately form the dataset used for model training and validation. Table 1 provides descriptions of the collected elevator data.

Table 1 Elevator Data Collection Description

Serial Number	Variable Name	Variable Type	Sampling Cycle	Remarks	Installation Location
1	Door Status	Discrete Variable	1s	800: Door Open, 700: Door Closed	Camera, installed on the elevator car side
2	Floor Position	Discrete Variable	1s		Calculated by the barometer installed in the elevator car
3	Leveling Status	Discrete Variable	1s	600: At Level, 500: Not at Level	Determined by photoelectric sensor at reference floors; determined by barometric pressure and whether movement speed is zero at non-reference floors
4	Travel Direction	Discrete Variable	1s	200: Up, 150: Stop, 100: Down	Determined by acceleration
5	Occupancy Status	Discrete Variable	1s	400: Occupied, 300: Unoccupied	Infrared sensor, installed inside the machine room near the control cabinet
6	Car Acceleration	Continuous Variable	10ms		Gyroscope, typically installed on the side of the elevator car, near the camera

Taken together, the elevator test bench built in this study and the data acquisition methods employed provide a solid foundation for developing elevator fault-diagnosis models. Through rigorous signal acquisition and data-processing procedures, the accuracy and reliability of the data were ensured, laying the groundwork for subsequent research. The time-series signals of ten types of elevator operational data and their corresponding labels are shown in Figure 5; these data reflect the elevator's states and behavioral patterns. By analyzing the acceleration data, one can determine whether a fault has occurred and identify its cause; therefore, vertical acceleration data are of particular importance for elevator fault diagnosis.

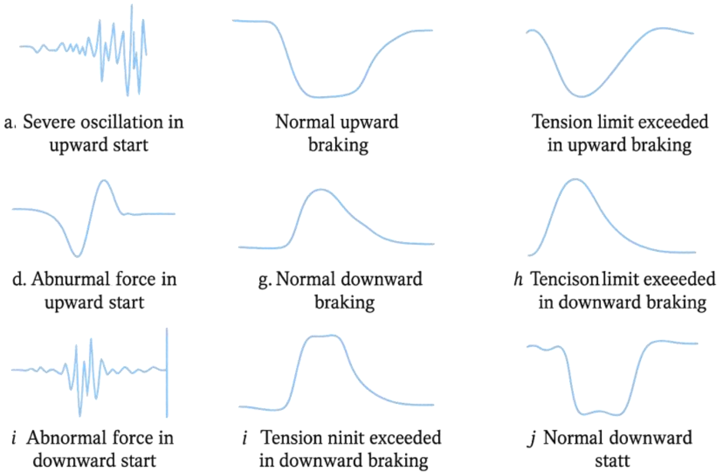


Figure 5 Elevator Fault Timing Signal

The figure above shows the time-series signals and labels corresponding to six fault types and four normal-operation conditions of the elevator. The plot indicates that time-series data for different fault types in the same travel direction differ markedly, while some operational data across opposite travel directions are highly similar—for example, normal descent start and normal ascent braking; normal descent braking and normal ascent start; and abnormal starting torque during ascent and abnormal starting torque during descent.

5.2 Dataset Construction

The quality of the dataset directly impacts model training effectiveness and diagnostic accuracy. In this study, dataset construction primarily involves collecting source domain data, acquiring target domain few-shot samples, data labeling, and preprocessing. First, the source domain dataset is built based on abundant data collected from the elevator testbed, which simulates real elevator operating environments and is equipped with various sensors to gather data under different states, including vibration, temperature, and current signals. These data are acquired in real-time via sensors and stored in a database. The source domain dataset covers multiple states such as normal operation, minor faults, and severe faults, providing comprehensive baseline data for model training. Acquiring the target domain few-shot dataset is more critical due to the typical scarcity of real-world fault data. This study obtained corresponding fault data by simulating specific fault patterns on the elevator testbed. Additionally, to enhance data diversity, elevator operation under different load and speed conditions was simulated, further enriching the target domain dataset. Data labeling is a crucial phase in dataset construction. Experienced elevator maintenance engineers were invited to annotate the collected data. The labeling process includes classifying normal and various fault states in the data, along with detailed annotations for specific fault types, ensuring accuracy and consistency. Preprocessing steps involve data cleaning, normalization, and feature extraction. Data cleaning removes outliers and noise to ensure data quality. Normalization scales feature values within the dataset to the same range, aiding numerical stability during model training. Feature extraction is the most critical step in preprocessing, where this study employs a multi-scale convolutional network to automatically extract effective features from the data. To verify dataset quality and applicability, statistical analysis was conducted. Statistics indicate the source domain dataset contains over 10,000 samples covering multiple elevator operating states, while the target domain few-shot dataset includes 500 representative fault samples. Comparative experiments reveal that dataset construction quality significantly influences model diagnostic performance. In summary, the constructed dataset not only encompasses rich source domain data but also emphasizes the acquisition of target domain few-shot samples. Through rigorous data labeling and preprocessing procedures, dataset quality is ensured, laying a solid foundation for subsequent model training and performance evaluation.

5.3 Experimental Protocol

To comprehensively evaluate the performance and applicability of the proposed model, this study designs a detailed experimental protocol including comparative experiments, ablation studies, and cross-domain scenario configurations. The following elaborates on the design rationale and implementation details of the experimental protocol. First, comparative experiments aim to validate the superiority of the proposed model in fault diagnosis tasks by comparing its performance with other advanced methods. For this purpose, multiple baseline models are selected, including traditional machine learning methods, deep learning approaches, and state-of-the-art few-shot learning techniques. All models are trained and evaluated on identical training and test sets to ensure fair comparison. Second, ablation studies are designed to investigate the specific contributions of key modules to model performance. By sequentially removing critical components such as the multi-scale convolutional structure in the feature extraction network, attention mechanisms, and the domain adaptation module, the impact of these components on overall model performance is assessed. Additionally, sensitivity to hyperparameter variations is analyzed through parameter adjustments. Cross-domain scenario configuration constitutes another essential part of the experimental protocol, aiming to examine the model's generalization capability in real-world applications. Specifically, elevators from different manufacturers, models, and service years are selected as data sources to construct multiple source domain datasets and target domain few-shot datasets. Cross-domain adaptability and diagnostic accuracy are evaluated through transfer experiments between different source and target domains. During experimentation, all datasets undergo rigorous preprocessing including data cleaning, normalization, and labeling. Furthermore, to simulate few-shot scenarios in practical applications, the sample size of target domain datasets is intentionally limited to a small range. Through this experimental protocol, the study aims to comprehensively evaluate the proposed model's performance from multiple perspectives, providing an effective solution for few-shot learning problems in elevator fault diagnosis. Experimental results will demonstrate not only the model's advantages in diagnostic accuracy but also its adaptability and stability across different scenarios.

6 EXPERIMENTAL RESULTS AND ANALYSIS

6.1 Diagnostic Performance Evaluation

Cross-domain transfer effectiveness serves as a key metric for evaluating the performance of domain-adaptive fault diagnosis models. This study conducts an in-depth analysis of model transferability across different domains by comparing the diagnostic performance of source domain-trained models and domain-adaptive models on target domain few-shot datasets. Statistics reveal that the overall accuracy of the source domain model without domain adaptation on the target domain dataset is significantly lower than that of the domain-adaptive model. Specifically, the average accuracy of the source domain model on the target domain dataset is 65%, while the domain-adaptive model achieves 85%. This result indicates that the domain adaptation strategy effectively enhances the model's generalization capability in the target domain.Further analysis using confusion matrices elucidates the diagnostic performance of the domain-adaptive model on specific categories. The source domain model exhibits significant misdiagnosis and missed diagnosis for certain fault types. For instance, the misdiagnosis rate for outer race bearing faults reaches 30% in the source domain model. After domain adaptation, the misdiagnosis rate for this fault type decreases to 5%, and the missed diagnosis rate drops to 10%. This demonstrates that the domain adaptation strategy significantly improves the model's recognition accuracy for specific fault categories.Regarding cross-domain transfer effectiveness, multiple comparative experiments were designed. Results show that the domain-adaptive model outperforms the source domain model across different target domain datasets. For example, on target domain dataset A, the domain-adaptive model's accuracy is 20% higher than the source domain model; on target domain dataset B, this gap is 15%. This indicates that the domain-adaptive model possesses strong cross-domain transferability, effectively improving fault diagnosis performance in various scenarios.Notably, the performance of the domain-adaptive model varies across different fault types and severity levels. The most significant performance improvement is observed for minor faults, while the enhancement is relatively smaller for severe faults. This phenomenon suggests that the model's capability for feature extraction and recognition differs under varying fault severities.Furthermore, this study analyzes the impact of different domain adaptation strategies on diagnostic performance. Experimental results show that both marginal distribution alignment and conditional distribution alignment strategies significantly improve model accuracy in the target domain. However, the dynamic adversarial training strategy occasionally leads to performance degradation, indicating that the selection and optimization of domain adaptation strategies are crucial for enhancing cross-domain transfer performance.In summary, the proposed domain-adaptive few-shot fault diagnosis model demonstrates significant advantages in cross-domain transfer effectiveness. Comparative experiments and confusion matrix analysis confirm the effectiveness of the domain adaptation strategy in improving model generalization and diagnostic accuracy. However, performance variations across different fault types and severities remain, necessitating further optimization of model architecture and domain adaptation strategies in future research.

6.2 Ablation Study Results

This study investigates the impact of individual components on the final diagnostic performance through carefully designed ablation experiments. By progressively removing key modules, we evaluate their respective contributions to overall performance. The following discusses hyperparameter sensitivity analysis in detail.First, the influence of different learning rates on model performance is examined. As a critical parameter in deep learning model training, the learning rate determines the magnitude of weight updates. Experimental results indicate that when the learning rate is set too low, model convergence is slow, and diagnostic accuracy improvement is insignificant. Conversely, an excessively high learning rate causes model oscillation, leading to unstable accuracy. Through repeated experiments, we find that a learning rate of 0.001 achieves high diagnostic accuracy while ensuring convergence speed in the current model.Second, the impact of different batch sizes on model performance is analyzed. Batch size determines the number of samples used in each training iteration, affecting model generalization capability and computational efficiency. Experiments show that while smaller batch sizes reduce memory consumption, they result in lower diagnostic accuracy. Larger batch sizes improve accuracy but require more computational resources, with diminishing returns beyond a certain threshold. A batch size of 64 achieves an optimal balance between diagnostic performance and computational efficiency.Next, the effect of different numbers of training iterations on model performance is explored. The number of iterations is another important hyperparameter determining training sufficiency. Experiments reveal that diagnostic accuracy gradually improves with increasing iterations, but the rate of improvement diminishes beyond a certain point while computational costs rise significantly. Therefore, 50 iterations are set in the current model to achieve satisfactory diagnostic accuracy.Additionally, the impact of different numbers of attention heads in the attention mechanism on model performance is investigated. The number of attention heads determines the granularity of the model's focus on input data. Results show that increasing the number of heads enhances the model's ability to parse input data, improving diagnostic accuracy. However, excessive heads increase model complexity and computational cost without significant accuracy gains. Thus, setting the number of heads to 4 maintains performance while avoiding excessive computational complexity.Through this hyperparameter sensitivity analysis, we demonstrate that different hyperparameter settings significantly affect model performance. Appropriate selection and adjustment of these parameters can effectively enhance diagnostic accuracy and generalization capability. These experimental results also validate the rationality and effectiveness of the proposed model architecture.

6.3 Visualization Analysis

Migration path tracking, as a crucial component of visualization analysis, provides an intuitive perspective for understanding the decision-making mechanism of the model during domain adaptation. Through a series of visualization techniques, this study demonstrates how the model learns from source domain data and adapts to target domain few-shot data, thereby improving fault diagnosis accuracy.First, feature distribution visualization reveals differences in feature representations between domains processed by the feature extraction network. Scatter plots in the feature space show significant distribution differences between source and target domain features. Statistics indicate that source domain features are relatively concentrated, while target domain features are dispersed with some overlap. These distribution differences provide a basis for designing domain adaptation strategies.Furthermore, attention heatmaps demonstrate the model's focus on key information during feature extraction. Taking vibration signals in elevator fault diagnosis as an example, the model extracts time-series features through a multi-scale convolutional structure and weights key frequency components using an attention mechanism. These visualizations intuitively reflect the model's focus on fault characteristics, providing important clues for understanding its decision logic.

In the visualization analysis of the domain adaptation module, the effects of marginal and conditional distribution alignment are clearly demonstrated. Comparing feature distributions before and after domain adaptation reveals that the distributions become more similar after alignment processing. This distribution convergence demonstrates the effectiveness of the domain adaptation strategy, ensuring the model's generalization capability in the target domain.Visualization tracking of the dynamic adversarial training strategy further reveals the model's adjustment strategies during adversarial processes. Through adversarial training, the model gradually learns differences between source and target domains and dynamically adjusts feature representations to reduce domain discrepancy. Tracking results show that as training progresses, decision boundaries in the feature space become clearer and better encompass target domain few-shot data.Moreover, visualization analysis of few-shot learning strategies highlights the roles of the meta-learning framework, data augmentation and synthesis, and prototype network optimization in improving model performance. Comparing feature distributions and attention heatmaps under different strategies clearly shows the significant contribution of few-shot learning strategies in enhancing model adaptability and diagnostic accuracy under few-shot conditions.In summary, visualization analysis not only provides intuitive evidence for the research but also deepens our understanding of the model's decision-making mechanism. This understanding helps identify potential weaknesses and guides future optimization. For example, by observing feature distribution visualizations, we can explore new feature extraction methods to improve cross-domain transferability. Simultaneously, attention heatmaps and migration path tracking provide important support for model interpretability, enhancing reliability and acceptability in engineering applications.

6.4 Discussion

Regarding engineering deployment feasibility, the proposed domain-adaptive few-shot fault diagnosis model demonstrates significant application potential. First, its performance in accuracy, confusion matrix analysis, and cross-domain transfer effectiveness provides strong technical support for practical elevator fault diagnosis applications. Second, ablation experiment results show that each module significantly contributes to overall performance, confirming the rationality and effectiveness of the model design.Feature distribution visualization analysis shows that data processed by the feature extraction network exhibits clearer feature distributions, facilitating improved fault diagnosis accuracy. Attention heatmaps further reveal key regions during fault feature identification, providing intuitive evidence for understanding the model's working mechanism. Additionally, migration path tracking demonstrates how the model dynamically adjusts feature learning strategies to adapt to target domain few-shot data during cross-domain adaptation. However, challenges remain for engineering deployment. First, the model's strong data dependency requires ensuring data quality and sufficiency in practical applications. Second, computational complexity is another consideration, particularly in scenarios with high real-time requirements, where balancing model complexity and diagnostic efficiency is crucial.Compared to existing methods, the proposed model demonstrates clear advantages in handling few-shot problems. Traditional methods often require large amounts of training data, whereas this model effectively utilizes limited data through domain adaptation and meta-learning strategies, improving generalization under few-shot conditions.Regarding engineering deployment feasibility, the model has been preliminarily validated on an elevator testbed, demonstrating practical value. However, achieving large-scale engineering application requires addressing several issues: First, model stability and robustness need testing on more elevator types and complex environments. Second, real-time performance requires algorithm and hardware optimization to meet real-time fault diagnosis needs. Third, interpretability requires further research to enhance decision transparency for engineer understanding and acceptance.In summary, the proposed domain-adaptive few-shot fault diagnosis model achieves significant theoretical and technical progress, providing a new solution for elevator fault diagnosis. Despite deployment challenges, further research and optimization could enable practical engineering applications. Future work will focus on multi-source domain transfer extension, online adaptive updates, and edge computing deployment to further enhance performance and feasibility in practical applications.

7 CONCLUSION

This study addresses the few-shot problem in elevator fault diagnosis by proposing a domain-adaptive few-shot fault diagnosis model. By introducing transfer learning and domain adaptation techniques and combining marginal-distribution and conditional-distribution alignment strategies, the model effectively mitigates the challenges of training with limited samples. Simultaneously, a feature-extraction network that fuses multi-scale convolution and attention mechanisms is designed and integrated within a meta-learning framework, substantially improving diagnostic accuracy and cross-domain generalization, thus providing theoretical and experimental support for engineering deployment. In terms of innovation, this work is the first in elevator fault diagnosis to introduce a domain-adaptation module that accounts for both marginal and conditional distribution alignment and to adopt a dynamic adversarial training strategy to reduce inter-domain discrepancies. An ensemble few-shot learning strategy combining meta-learning, data augmentation, and a prototypical network is constructed to enhance learning capacity under limited samples. Nevertheless, the study has limitations, including strong dependence on data quality, relatively high computational complexity, difficulty of deployment on resource-constrained devices, limited ability to recognize rare faults, and lack of real-time online adaptation. Future research will focus on three directions: multi-source domain transfer expansion, online adaptive updating, and edge-computing deployment. By integrating knowledge from multiple source domains, developing online learning algorithms, and advancing model lightweighting and compression techniques, we aim to build a more robust, real-time, and efficient intelligent diagnostic system suitable for complex and variable real-world engineering environments.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Qing Guangwei, Liu Xiaofan. Research on Prediction Method of City Elevator Entrapment Fault Causes Based on Machine Learning. Internet of Things Technologies, 2022, 12(10): 55–58.

[2] Qi Yongsheng, Shan Chengcheng, Gao Shengli, et al. Fault diagnosis strategy for wind turbine bearing based on AEWT-KELM. Acta Energiae Solaris Sinica, 2022, 43(8): 281–291.

[3] Zhao Wenqiang, Zhou Jun, Wang Zhengwei, et al. Fault diagnosis method of synchronous condenser used in UHV transmission system. Science Technology and Engineering, 2024, 24(9): 3683–3690.

[4] Jiang Wanlu, Zhao Yan, Li Zhenbao. Fault diagnosis method for rotating machinery based on multi-model stacking ensemble learning. Chinese Hydraulics & Pneumatics, 2023, 47(4): 46–58.

[5] Zhao Bochao, Ma Jiajun, Cui Lei, et al. Photovoltaic anomaly detection based on improved VMD-XGBoost-BILSTM hybrid model. Computer Engineering, 2024, 50(3): 306–316.

[6] Liu Y, Jiang H, Yao R, et al. Counterfactual-augmented few shot contrastive learning for machinery intelligent fault diagnosis with limited samples. Mechanical Systems and Signal Processing, 2024, 216: 111507.

[7] Chen Yuanqiong, Meng Yujia, Li Zhihao. Research on Distributed Fault Diagnosis System Based on Machine Learning. Computer Knowledge and Technology, 2024, 20(03): 22–24.

[8] Song Zhangting, Liu Yang, Guo Liang. Research on Causes and Solutions of Vertical Elevator Faults. China Plant Engineering, 2023(20): 170–173.

[9] Zhu Wenhua, Zuo Yi. Design of Elevator Control System Based on S7-1500PLC. Machine Building & Automation, 2023, 52(06): 182–186.

[10] Bi Lilong. Research on Fault Detection Method of Elevator Brake Based on Machine Vision. China Machinery, 2023(10): 116–119.

[11] Li Guodong. Discussion on Fault Diagnosis and Maintenance of Elevator Electrical Control System. Modern Industrial Economy and Informatization, 2023, 13(02): 246–248.

[12] Wu Cong, Li Mengnan, Li Kun. Elevator Bearing Fault Detection Based on PBO and CNN. Information Technology, 2023, 47(04): 73–78.

[13] Niu Dapeng, Guo Lei, Zhang Weiwei, et al. Operation performance evaluation of elevators based on condition monitoring and combination weighting method. Measurement, 2022, 194(1): 13–17.

[14] Ali Murad, Din Zakiud, Solomin Evgeny, et al. Open switch fault diagnosis of cascade H-bridge multilevel inverter in distributed power generators by machine learning algorithms. Energy Reports, 2021, 7(1): 13–20.

[15] Chen Zhiyu. An Elevator Car Vibration Fault Diagnosis Based on Genetic Optimization RBF Neural Network. China Science and Technology Information, 2023(10): 110–115.

[16] Feng Yongming. Research on Data Mining of Elevator Safety Big Data Based on Hadoop. Xi'an: Xi'an University of Science and Technology, 2023.

[17] Meng Lin, Wang Xiaoyang, Guo Qingliang. Analysis of Related Technical Issues in Elevator Fault Detection. Engineering Technology Research, 2020, 5(7): 62–63.

[18] Zhang Zhanyi, Zhang Baoquan, Wang Zhouli, et al. Data Augmentation Optimization for Multi-Tea CNN Image Recognition and Quantitative Evaluation of Class Activation Mapping. Journal of Tea Science, 2023, 43(3): 411–423.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3099

KEY NODE IDENTIFICATION ALGORITHM BASED ON LOCAL SEMI GLOBAL TRIANGLE CALCULATION

HanYi Yang[1*], YiJia He[2]
[1]College of Cyber Security, Tarim University, Alar 843300, Xinjiang, China.
[2]College of Foreign Languages, Tarim University, Alar 843300, Xinjiang, China.
Corresponding Author: HanYi Yang, Email: xzmu_yhy@qq.com

Abstract: Aiming at the challenges of low identification accuracy and slow computation time in existing key node identification algorithms for complex networks, the paper proposes a key node identification algorithm based on local semi global triangular computation (LSTC). First, inspired by the structural stability of triangles in the physical world, the triangular patterns of nodes in complex networks and their importance are defined. Second, drawing on the third-order partition theory which highlights strong connections between a node and its third-order neighbors, the algorithm incorporates the influence of a node's local third-order neighbors when evaluating its importance. To validate the experimental performance of the proposed algorithm, the LSTC algorithm is compared with eight other algorithms of the same type using both the Susceptible–Infected–Recovered (SIR) model and the Linear Threshold (LT) model. Experimental results demonstrate that the proposed algorithm achieves the highest overall performance.

Keywords: Complex network; Influential spreaders; Spreading ability; SIR epidemic model

1 INTRODUCTION

Network communication is a natural phenomenon in many fields of life[1], such as pandemic, disease transmission, information transmission, etc. We must adjust or maximize the communication process according to social interests and requirements. Influential communicators play a key role in optimizing or managing the impact of the communication process. The most influential extender is an important node in the network, which acts as the maximizer or controller of the extender process. In the real world, influential communicators have many applications in various fields, such as spreading information on social networks, and controlling rumors or epidemics in the system. In order to identify influential communicators from the network, researchers have proposed various centralized methods. The central approach is also used in other areas. For example, measure the social impact capacity of scientists in the cooperative network to determine the important economic pillars in the economic network, investigate the impact and timeliness of journals in the scientific citation network, and find communities. In this article, we focus on identifying influential communicators from the network. The schematic diagram of a complex network is shown in Figure 1:

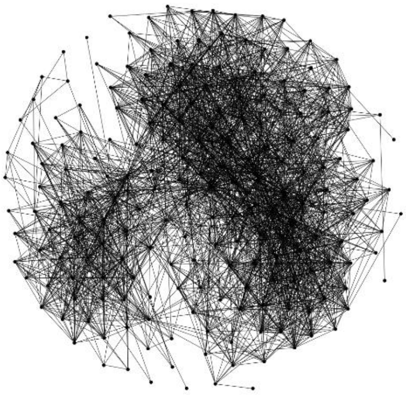


Figure 1 Complex Network

The identification of key nodes in complex networks is reflected in various fields. In the public health epidemic [2], identifying key individuals will help optimize vaccination strategies to contain the outbreak of the epidemic [3]. In social media and digital marketing, identifying well connected users can maximize the scope and impact of information activities and spread content. Similarly, in network security and error information control, locating the central node allows to contain rumors or malicious attacks in the network system. In addition, the centrality based approach also helps to quantify the scientific impact in academic cooperation networks, identify systemically important entities in financial networks, evaluate the reputation of journals in citation networks, and detect community structures in social and biological networks.

Although researchers have designed various centrality measures to sort and select the importance of complex network nodes, many measures are limited by their dependence on specific structural attributes or computational constraints. In order to solve the problem, this paper pays special attention to the identification of influential communicators in complex networks. We systematically reviewed and classified the existing centralized methods, emphasized their advantages and limitations, and stimulated the demand for more robust and efficient methods that can adapt to different network topologies and dynamic contexts.

Among various selection algorithms, the most widely used metric is usually based on the topological position of nodes or some aspects of extended dynamics, such as random walking, shortest path distance and information content. These centrality methods can be roughly divided into four types: local centrality, global centrality, semi global centrality and mixed centrality.

2 RELATED WORK

2.1 Basic Algorithm

The key node identification algorithm can be applied to solve various real-world problems. For example, in a water supply network, pollutants in the entire system can be monitored by identifying key locations and deploying sensors at these points. The core challenge of such algorithms is how to effectively and accurately identify the most influential nodes in the network. In recent years, researchers have proposed a series of key node identification methods, which are summarized as follows:

(1) Degree

The most basic topological property in the network is the degree of the node. The algorithm based on degree centrality (Degree algorithm) [4] believes that the greater the degree of the node, the higher the influence in the network, that is, the more direct neighbors of the node, the more important the node is. The method for calculating node degree centrality is defined as:

$$DC_v = \frac{|\Gamma_v|}{N-1} \tag{1}$$

Where, Γ_v is the set of direct neighbors of a node v , $|\Gamma_v|$ is the module of the set of direct neighbors of a node, N is the total number of nodes in the network, and $N-1$ is the number of other nodes in the network except nodes.

(2) H-index

H-index [5] is a widely used measurement standard, which is used to evaluate the scientific research output of researchers and the influence of papers. Early measurement of academic influence usually relied only on isolated indicators, such as the total number of publications or the number of citations of each paper. However, the H index achieves a more balanced assessment by taking into account the number of authors' publications and the corresponding number of citations. This dual consideration enables a more accurate and robust reflection of a scholar's academic influence. H is defined as an operator, which only acts on the finite set of real numbers $(x_1,x_2,\cdots,x_n)$, so H-index is defined as:

$$y = H(x_1,x_2,\cdots,x_n) \tag{2}$$

Where, $y>0$ and y is the largest integer, the H operator makes that there are at least y elements in $(x_1,x_2,\cdots,x_n)$, and each element is not less than y . In other words, for a scholar with n papers, $(x_1,x_2,\cdots,x_n)$ is the number of citations of the papers, and $H(x_1,x_2,\cdots,x_n)$ is the H-index value of the scholar.

(3)K-shell

The K-shell algorithm was first proposed by Kitsak et al [6]. and used to find influential nodes in the network. The algorithm believes that the closer the node is to the center of the network, the greater its importance and influence. The specific process of K-shell algorithm is as follows:

(a) Count the degree of each node in the network and record the minimum degree $k_{\min}$.

(b) The k-shell decomposition process iteratively prunes nodes with degree $k_{\min}$ from the network. After each removal phase, the degrees of remaining nodes are updated, and the procedure repeats for the new graph structure. Each pruned set of nodes is assigned the current k-shell value. Nodes removed in later iterations—those persisting in more densely connected core regions—receive higher k-shell values, reflecting their greater structural centrality within the network.

(c) Repeat (a) and (b) until there are no connected nodes in the network, that is, each node has a k-shell value.

(4) ISK

Based on the K-shell algorithm, Wang et al[7]. introduced the ISK algorithm by incorporating the concept of information entropy—also referred to as Shannon entropy, which serves as a quantitative measure of information. The IKS algorithm utilizes the magnitude of a node's information entropy to differentiate nodes that share the same k-shell value, thereby enabling a more refined and accurate ranking of node influence within the same shell layer. First, the local importance of node i is defined:

$$I_i = k_i \Big/ \sum_{j=1}^{N} k_j \tag{3}$$

Where, k_i is the degree of the node, and j is the direct neighbor of the node i . Secondly, the method to calculate the node information entropy is as follows:

$$e_i = -\sum\nolimits_{j\in\Gamma(i)} I_j \cdot \ln I_j \tag{4}$$

(5) MCDE

Based on the idea of hierarchical division of nodes in K-shell algorithm, many scholars have proposed various improved algorithms based on K-shell. Among them, Sheikhahmadi et al. [8]. This paper introduces the MCDE algorithm, which integrates the degree, k-shell value and information entropy of nodes into the weighted comprehensive measure. This multidimensional approach aims to capture node impacts more comprehensively. It is worth noting that the calculation of information entropy in MCDE algorithm is slightly different from that used in IKS algorithm. The specific formula of information entropy in MCDE is defined as follows:

$$Entropy(v) = -\sum\nolimits_{i=0}^{Core_{\max}} (p_i \cdot \log_2 p_i) \tag{5}$$

Where, $p(i)$ is the ratio of the cumulative sum of the k-shell values of the direct neighbors of the node to the node degree, and the calculation formula is as follows:

$$p_i = \frac{\sum\nolimits_{j\in\Gamma(i)}^{N} \text{k-shell(j)}}{k(i)} \tag{6}$$

Based on the above considerations, the MCDE algorithm is formally defined as follows:

$$MCDE(v) = \alpha \cdot \text{k-shell}(v) + \beta \cdot k(v) + \gamma \cdot Entropy(v) \tag{7}$$

Where, α , β , and γ are respectively k-shell values, node degrees, and weights of information entropy.

(6) ECRM

Based on the principle that the node importance in clustering algorithm can be calculated through the interaction between direct neighbors, Zareie et al. [9]. ECRM algorithm is proposed based on clustering algorithm. This method goes beyond simply calculating the shared neighbors between nodes and their direct connections; It also integrates the structural commonalities between them. Specifically, the algorithm makes use of the similarity and correlation between the connection modes of the connecting node and its immediate neighbors.

(7) VoteRank

In 2016, Zhang et al [10]. pioneered the application of a voting strategy to identify influential nodes by proposing the VoteRank algorithm. This method initializes each node in the network with a certain voting capacity, allowing it to cast votes in favor of its direct neighbors. During each iteration, the node accumulating the highest number of votes is identified as the most influential node in that round. The detailed procedure of the VoteRank algorithm is outlined as follows:

(a) Each node in the network is assigned a tuple (voting score, voting capacity), that is, (s_u, va_u) , and initialized to $(s_u, va_u) = (0,1)$.

(b) It is stipulated that each node can vote for its immediate neighbors according to its own voting capacity. The voting score of a node is the sum of the voting capacity of each neighbor. The voting score is calculated as:

$$s_u = \sum\nolimits_{v\in\Gamma_u} va_v \tag{8}$$

(c) The node with the highest voting score in the current round is selected as the influential node. Once selected, this node is excluded from participating in any subsequent voting rounds.

(d) The voting capacity of the direct neighbors of the selected influential node is reduced by a factor equal to the reciprocal of the network's average degree $< k >$, as defined by the following expression:

$$va_u = va_u - \frac{1}{< k >} \tag{9}$$

(e) Repeat step (b), (c) and (d) until the specified L influence nodes are selected, where $L < N$.

(8) EnRenew

Based on the VoteRank algorithm, Guo et al. [11]. The EnRenew algorithm was proposed, which combines node information entropy to enhance the original method. This method effectively addresses the lack of discrimination in initial voting capacity allocation and subsequent voting decay processes in VoteRank. This algorithm utilizes node information entropy to initialize different voting abilities and weakening factors for different nodes, thereby achieving a more refined and adaptive node influence evaluation mechanism.

The EnRenew algorithm has been innovated in the initialization and weakening stages of the voting process. It uses information entropy to determine the initial voting ability and applies different attenuation factors during the weakening stage. Compared with the VoteRank algorithm that uses unified voting parameters, this improved method has achieved significant improvements in algorithm performance and recognition accuracy.

2.2 Analysis Summary

The above eight key node recognition algorithms have limitations in accuracy and computational efficiency, but the VoteRank algorithm provides an innovative solution by introducing some methods from society into complex networks. Similarly, this article also draws on the stability of triangular structures in the real world to construct a node neighbor triangular pattern in complex networks. At the same time, it draws on the strong connection between nodes and their three boundary neighbors in the theory of three-degree separation, and comprehensively considers the local importance of nodes. The next section will provide a detailed introduction to our method.

3 PROPOSED ALGORITHM

3.1 Basic Concept

In large-scale real complex networks, a small number of key nodes often have a significant impact on the overall system behavior. Therefore, accurately identifying key nodes in a network plays an important role in the overall inference of complex networks. In order to improve the accuracy and efficiency of key node recognition, this paper proposes a key node recognition algorithm based on local semi global triangular model calculation. Firstly, inspired by the inherent stability of triangular structures in the real world, the concept of triangular patterns was introduced into complex networks, and the triangles formed between nodes and their neighbors were calculated. Secondly, based on the strong connection between nodes and their third-order neighbors in the theory of three degree segmentation, the importance of nodes is fully considered by taking into account their third-order neighborhoods.

3.2 Triangle Mode Calculation

The size of network density is closely related to the number of triangles in the network. The more triangles the network has, the greater the network density; The smaller the number of triangles, the smaller the network density. For a complex network $G=(V,E)$ with no right and no direction, its network density is calculated as follows:

$$D=\frac{E}{|V|*(|V|-1)/2} \tag{10}$$

Where E is all the edges in the network, and V is the collection of nodes in the network.

For a complete network or a completely dense network, the density is $D=1$. In a completely dense network, the total number of triangles is calculated as follows:

$$^{|V|}C_{R=3}=\frac{|V|!}{3!(|V|-3)!} \tag{11}$$

Where, $R=3$ indicates that the triangle vertex contains three nodes. The specific calculation is shown in Figure 2:

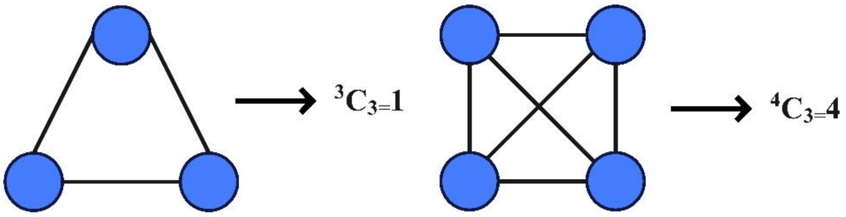


Figure 2 Triangle Calculation

Figure 2 illustrates the triangular counting principle in fully connected networks. The left panel depicts a 3-node fully connected network, which contains exactly one triangle. The right panel shows a 4-node fully connected network (K4), where the total number of distinct triangles amounts to four.

The four node fully connected network in the right figure is gradually removed. The calculation process of network density and triangle number in the removal process is shown in Figure 3:

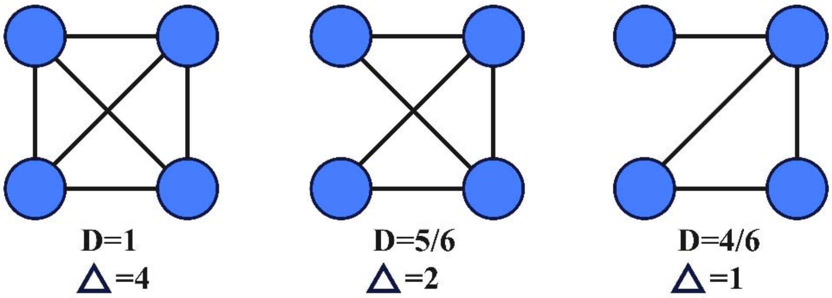


Figure 3 Gradually Remove Network Edges

Figure 3 demonstrates the progressive edge removal process applied to an initially fully connected four-node network. In the middle graph, where one edge has been removed, the network density measures 0.833 with three remaining triangles. The right graph, with two edges removed, shows a further reduced density of 0.66 and only one preserved triangle. These results clearly indicate a positive correlation between network density and the number of triangular structures. Therefore, the importance of a node can be gauged by the number of triangles it forms with its neighboring nodes.

Therefore, the pseudo code of the algorithm to form a triangle through nodes and their neighbors is as follows:

Algorithm 1: Triangle Calculation

```
Input:   = ( , ), Set: TempList1[], TempList2[], TempList3[], TempList4[]
Output: Triangle List Final [i, number of Triangle]
for  ∈  do
    number of Triangle=0;
    for  ∈  h  ( ) do
        for k ∈  h  (j) do
            TempList1= getNeighbors(i);
            TempList2= getNeighbors(j);
            TempList3= getNeighbors(k);
            TempList4 = intersection(TempList1, TempList2, TempList3);
            number of Triangle += Size of TempList4;
    Update Triangle List Final [ , number of Triangle⁄2];
end;
```

3.3 Voting Score Calculation

According to the number of third-order neighborhood triangles of nodes calculated in the upper part, the semi global triangular centrality of nodes will continue to be calculated as follows. According to the characteristic that the connection between a node and its third-order neighbor node is a strong connection in the third-order partition theory, when calculating the triangle centrality, consider the influence of the third-order neighbor node according to different weights from large to small. The specific calculation is as follows：

$$st_c(m) = \sum_{n \in N_m} \triangle_n / 2^d \qquad (12)$$

Where, $\triangle_n$ is the number of triangles generated from node n , and N_m is the neighborhood set including node m . The node set consists of node m , the nearest neighbor of node m , and the next nearest neighbor. d represents the distance between node m and neighborhood set N_m . Here, $d=1$ represents the nearest neighbor node of node m, $d=2$ represents the second order neighbor node of node m, and $d=3$ represents the third order neighbor node of node m. According to the characteristics of strong connection between nodes and their third-order neighbors in the third-order partition theory, consider the third-order neighbors of nodes, because the third-order neighbors of nodes are the best choice to balance the spread spectrum cost and performance. If more neighbor steps are considered, the ranking effect will not be significantly improved, or even decline. The specific pseudo code is as follows:

Algorithm 2: Local Semi-Global Triangular Centrality

```
Input:   = ( , ), Triangle List Final [i, number of Triangle]
Output: Rank [m, LSTC(m)]
for  ∈  do
    rank=0;
    rank= Triangle List Final [ ]/2^0;
    for  ∈  h  ( ) do
        rank+= Triangle List Final [ ]/2^1;
        for  ∈  h  ( ) do
            rank+= Triangle List Final [ ]/2^2;
    Update Rank [m, rank];
end;
```

4 EXPERIMENT AND ANALYSIS

4.1 Datasets and Comparison Algorithms

In order to verify the authenticity and effectiveness of the proposed algorithm LSTC, the following performance comparison tests will be carried out. The eight algorithms of the same type are: ECRM、EnRenew、MCDE、Degree、H-index、ISK、K-shell、VoteRank. These algorithms are mentioned above. The six real network data sets required for the experiment are shown in Table 1:

Table 1 Datasets

NetWork	N	M	<k>
Hamster	2426	16631	13.71
NetSci	379	914	4.82
PGP	10680	24316	4.55
Power	4941	6594	2.66
USAir2010	1574	17215	21.87
Yeast	2224	6609	5.94

Where, N is the total number of nodes in the complex network, M is the total number of sides in the complex network, and $<k>$ is the average degree of the network. The relationship between the three network topological properties is $N\cdot <k> = 2\cdot M$.

4.2 Experimental Model

To accurately evaluate the quality of nodes selected by key node identification algorithms, researchers have developed propagation models. To date, the most widely used models include the Independent Cascade Model (IC), the Infectious Disease Model (SIR), and the Linear Threshold Model (LT), each with its own evaluation metrics. In order to comprehensively and accurately measure the performance of the algorithm proposed in this paper, the Infectious Disease Model and the Linear Threshold Model are selected here. The indicators associated with these models will be described in detail in the following sections.

4.2.1 SIR

The infectious disease model, commonly known as the SIR model [12], is a classic epidemiological framework. Originally developed to quantify the scope and efficiency of virus transmission within a population, it has now been widely used to evaluate the quality of influential nodes identified by impact maximization algorithms [13]. In this case, the set of nodes that initiate earlier and faster propagation is considered to have higher quality. This model divides nodes into three different states: susceptible (S), infected (I), and recovered (R). Lymph nodes in S state are healthy and prone to infection; State I indicates that the node has been infected and is contagious; State R indicates that the node has recovered and gained immunity. It is crucial that each node can only exist in one state at any given time. In addition, state transitions are usually irreversible within a single propagation cycle - specifically, the recovered node cannot become sensitive again, which means that a person is only infected with the virus once.

In the initial stage of simulation, all nodes in the network are set to a sensitive (S) state. Then, the key nodes identified by the algorithm are designated as the initial infection source and transformed into the infection (I) state. Subsequently, each infected node attempts to infect its susceptible neighbors with an infection probability μ , while each affected node can recover with a recovery probability ξ and enter a recovery (R) state. Once restored, immune lymph nodes can no longer be reinfected by any infected neighbors. The values of infection probability and recovery probability are crucial as they directly determine the final scale of transmission. A too high A can cause infection to saturate the network too quickly, thereby masking the relative influence of the initial seed nodes. On the contrary, a too small A may suppress the propagation process and even prevent high impact nodes from triggering meaningful cascades, making it difficult to distinguish their effects.

Based on the structural characteristics of the infectious disease model, two indicators to quantify the influence of nodes are proposed, namely, the infection scale $F(t)$ [14] and the final infection scale $F(t_c)$. Among them, the infection scale refers to the limit value of infected nodes in the network over time. The calculation of the infection scale is as follows:

$$F(t) = \frac{n_{I(t)} + n_{R(t)}}{N} \tag{13}$$

Where, N is the total number of nodes in the network, t is the time, $n_{I(t)}$ is the number of nodes in the network in state I at time t , $n_{R(t)}$ is the number of nodes in the network in state R at time t , where, $N = n_{S(t)} + n_{I(t)} + n_{R(t)}$ 。 It can be seen that the infection scale in the epidemic model represents the proportion of infected nodes and cured nodes infected with viruses in the total number of nodes.

The final infection scale of another quantitative indicator is defined as follows:

$$F(t_c) = \frac{n_{R(t_c)}}{N} \tag{14}$$

Where $n_{R(t_c)}$ is the limit value of the number of nodes in the network with R status over time. It can be seen that the final infection scale in the epidemic model represents the proportion of all infected nodes in the total number of nodes at the end of the virus transmission.

4.2.2 LT

Linear threshold model, also known as LT model [15], is another commonly used influence model. The model has two node states: active state and inactive state. A node in the active state has an activation probability θ , and a node in the inactive state has an activation threshold upper bound θ_{active} . At the initial stage, all nodes in the network are in the inactive state, and the node state selected by the influence maximization algorithm is in the active state. The active node attempts to activate its direct neighbor with the activation probability θ . The inactive node accumulates the activation probability θ brought by the direct neighbor. If the cumulative value exceeds the activation threshold upper bound, it is activated, and the node state is converted from the inactive state to the active state. The mathematical definition of state transformation is as follows:

$$\theta_{active} \geq \sum_{i \in \Gamma_v} \theta_i \tag{15}$$

4.3 Experimental Result

Before the experimental comparison based on the infectious disease model, the experimental parameters should be tested first. In order to accurately describe the relationship between infection probability and recovery probability in the infectious disease model, the infection rate $\beta = \mu / \xi$ is defined. The infection rate β is proportional to the infection probability ξ and inversely proportional to the recovery probability B. The infection rate will directly affect the scale of virus infection in the infectious disease model. When the infection rate is too high, the scale of virus infection in complex networks is too fast, which is not conducive to the analysis of node influence; When the infection rate is too small, the virus infection scale of the complex network is too slow, and the convergence time of the infection scale is too long, which is not conducive to observing changes.

For this reason, we first designed a comparative experiment between infection rate β and final infection scale $F(t_c)$ to find the infection rate most suitable for the spread of infection scale. The specific experiment is shown in Figure 4. It can be seen from Figure 4 that in Hamster, PGP and USAir2010, no matter what the infection rate is, the final infection scale of LSTC always reaches the maximum. In NetSci, when the infection rate is 0.9, the performance of LSTC is slightly lower than EnRenew. When the infection rate is 0.3, 0.6, 1.2, 1.5, 1.8, the performance of LSTC is higher than EnRenew. In Power, LSTC algorithm is only slightly lower than EnRenew when the infection rate is 0.3. In Yeast, when the infection rate is 0.3 and 0.9, the effect of LSTC algorithm is better than ISK, and in other cases, it is worse than ISK. Even so, LSTC still exceeds the other seven algorithms. It can be seen that when the infection rate is 0.9, the comprehensive performance is the best. When conducting the infection scale experiment, the infection rate is set to 0.9.

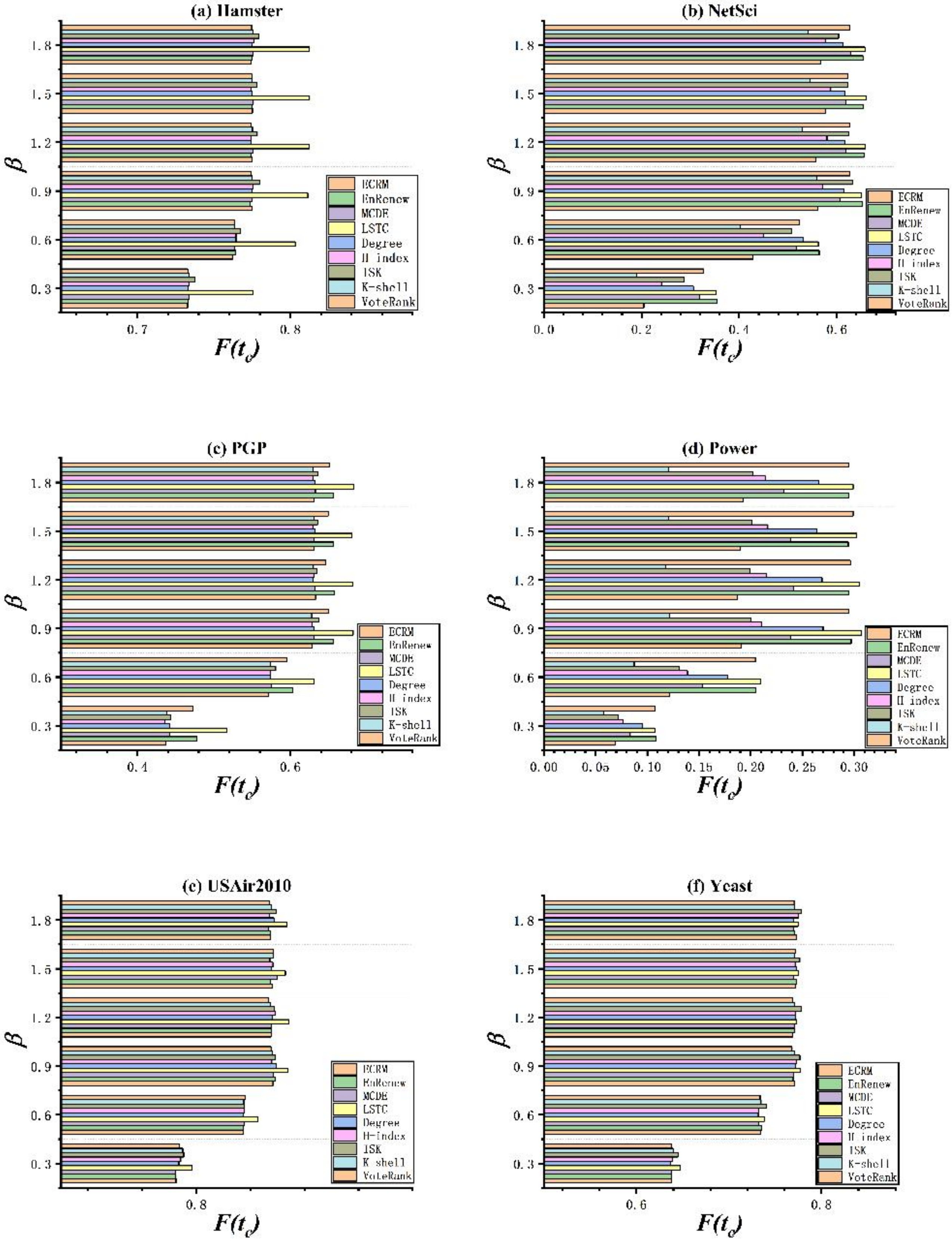


Figure 4 Experiment on Final Infection Scale and Infection Rate

(1) The comparative experiment of infection scale and time based on infectious disease model is shown in Figure 5:

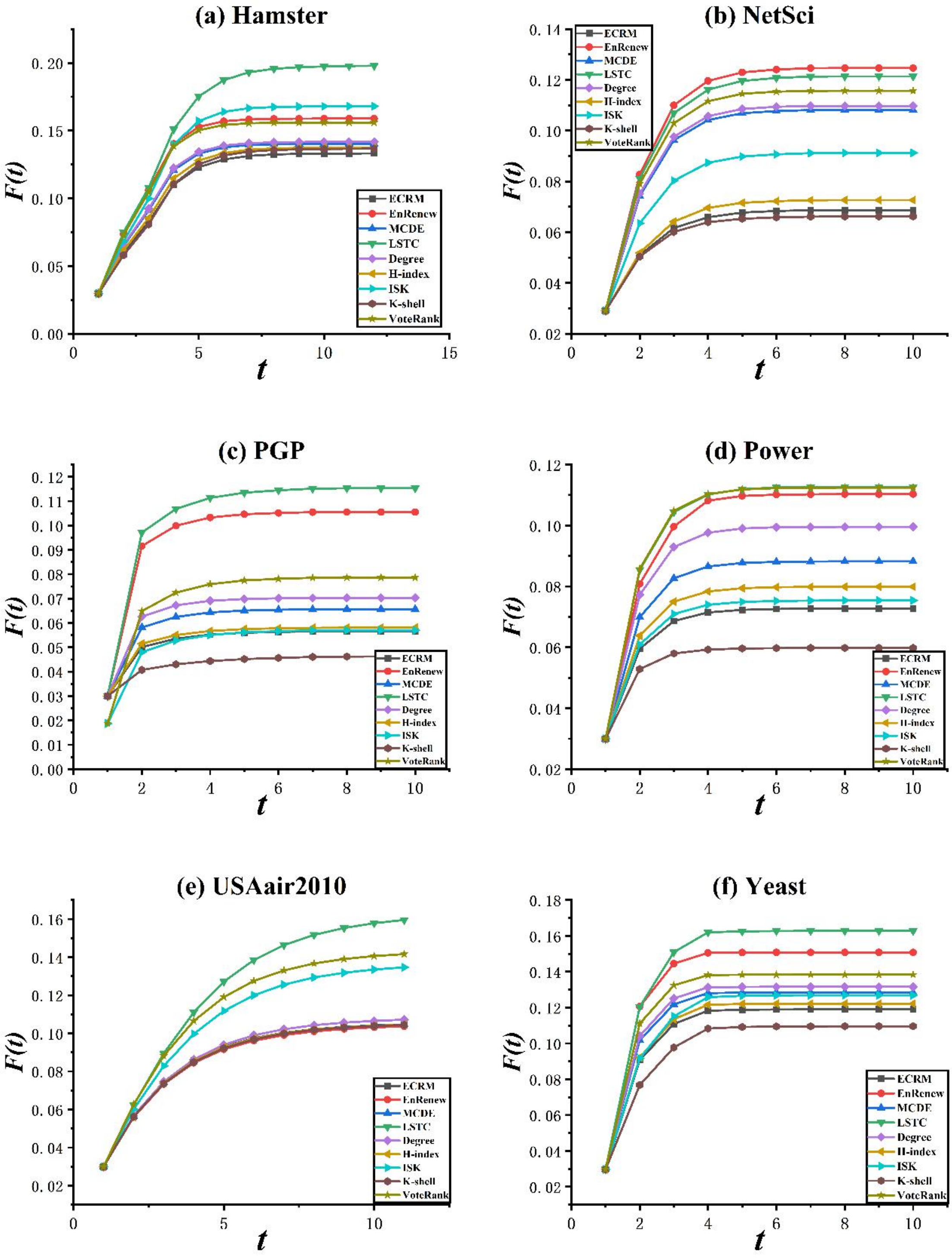


Figure 5 Experiment on Infection Scale and Time

It can be seen from Figure 5 that in the dataset NetSci, as time goes by, the infection scale of key nodes selected by LSTC algorithm is higher than that of ECRM, MCDE, Degree, H-index, ISK, K-shell and VoteRank algorithms. But it is slightly lower than EnRenew algorithm. Thus, the experimental performance of LSTC algorithm is the best.

(2) The comparison experiment between the number of active nodes based on linear threshold and the initial infection ratio is shown in Figure 6:

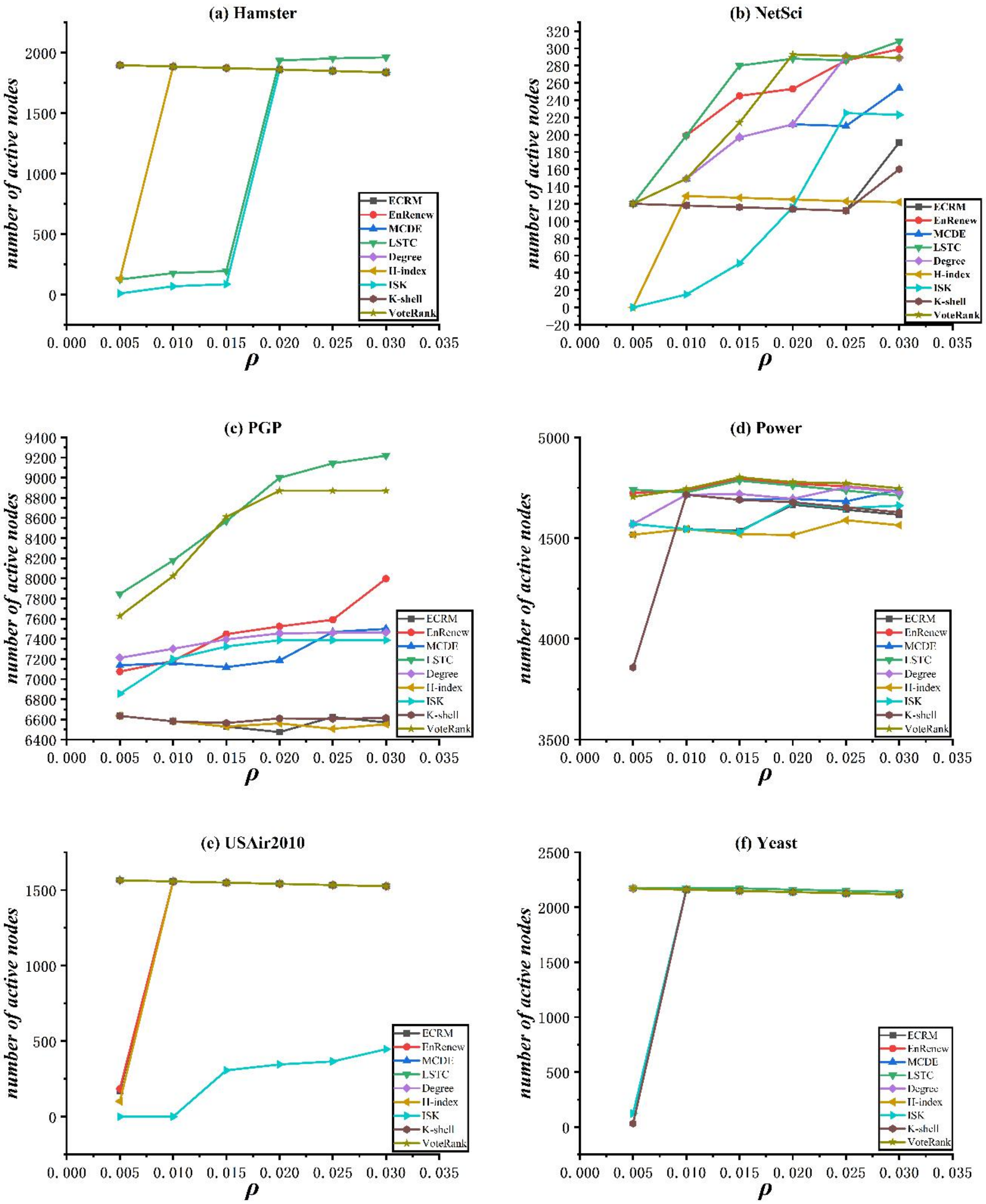


Figure 6 Experiment on the Number of Active Nodes and the Initial Infection Ratio

Where, ρ is the initial infection ratio, which is the proportion of nodes selected by the key node identification algorithm in the total network nodes. It can be seen from Figure 6 that in Hamster, when the initial infection ratio is greater than 0.02, the performance of LSTC algorithm is better than other algorithms, and when it is less than 0.02, it is only better than ISK algorithm. In NetSci, only when the initial infection ratio is 0.02 and 0.025, the performance of the algorithm is slightly lower than that of the VoteRank algorithm. In other cases, it is better than the other eight algorithms. In the dataset PGP, the performance of LSTC algorithm is basically higher than that of other algorithms, reaching the best. In Power, although the performance of this algorithm cannot reach the highest level, the overall trend is better than other algorithms. In the USAir2010 and Yeast data sets, LSTC performance is higher than or equal to other algorithms. In conclusion, LSTC algorithm has the best performance.

5 CONCLUSION

In order to improve the accuracy of key node identification in complex networks, this paper proposes the LSTC algorithm based on the three degree segmentation theory and triangular pattern calculation. This algorithm first introduces the real world stable triangular pattern into the complex network, and measures its importance by calculating the number of triangles formed by nodes and their neighbors. Secondly, referring to the characteristic of strong connection between nodes and their third-order neighbors in the three-dimensional partition theory, the number of triangles formed by nodes and nodes in their third-order local neighborhood is taken as the final calculation basis. By comparing eight algorithms of the same type on six real data sets: experiments based on infection scale and time of infectious disease model, experiments based on the number of active nodes and initial infection ratio of linear threshold model, LSTC algorithm has the best comprehensive performance.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This paper is supported by the Tarim University Euphrates Poplar Talent Introduction Scientific Research Foundation Project (TDZKSS202546).

REFERENCES

[1] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. Washington, USA, 2023: 137-146.
[2] Wu P, Pan L. Scalable influence blocking maximization in social networks under competitive independent cascade models. Computer Networks, 2017, 123: 38-50.
[3] Leskovec J, Adamic L A, Huberman B A. The dynamics of viral marketing. ACM Transactions on the Web (TWEB), 2007, 1(1): 5-es.
[4] Freeman L C. Centrality in social networks: Conceptual clarification. Social network: critical concepts in sociology. Londres: Routledge, 2002, 1: 238-263.
[5] Lü L, Zhou T, Zhang Q M, et al. The H-index of a network node and its relation to degree and coreness. Nature communications, 2016, 7(1): 10168.
[6] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks. Nature physics, 2010, 6(11): 888-893.
[7] Wang M, Li W, Guo Y, et al. Identifying influential spreaders in complex networks based on improved k-shell method. Physica A: Statistical Mechanics and its Applications, 2020, 554: 124229.
[8] Sheikhahmadi A, Nematbakhsh M A. Identification of multi-spreader users in social networks for viral marketing. Journal of Information Science, 2017, 43(3): 412-423.
[9] Zareie A, Sheikhahmadi A, Jalili M, et al. Finding influential nodes in social networks based on neighborhood correlation coefficient. Knowledge-based systems, 2020, 194: 105580.
[10] Zhang J X, Chen D B, Dong Q, et al. Identifying a set of influential spreaders in complex networks. Scientific reports, 2016, 6(1): 27823.
[11] Guo C, Yang L, Chen X, et al. Influential nodes identification in complex networks via information entropy. Entropy, 2020, 22(2): 242.
[12] Kermack W O, McKendrick A G. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 1927, 115(772): 700-721.
[13] Buscarino A, Fortuna L, Frasca M, et al. Disease spreading in populations of moving agents. Europhysics Letters, 2008, 82(3): 38002.
[14] Ma L, Ma C, Zhang H F, et al. Identifying influential spreaders in complex networks based on gravity formula. Physica A: Statistical Mechanics and its Applications, 2016, 451: 205-212.
[15] Granovetter M. Threshold models of collective behavior. American journal of sociology, 1978, 83(6): 1420-1443.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3100

A BI-DIRECTIONAL CASCADED RELATION EXTRACTION MODEL BASED ON DYNAMIC GATING MECHANISM: AN ENHANCED APPROACH TO OVERLAPPING TRIPLES

JiaBao Wang*, ZhiBin Guo
College of Cyber Security, Tarim University, Alar 843300, Xinjiang, China.
Corresponding Author: JiaBao Wang, Email: 2826698337@qq.com

Abstract: Relation extraction is a core task in natural language processing, aiming to identify semantic relationships between entities in unstructured text. Existing relation extraction methods face significant challenges when dealing with overlapping triples, especially in Entity Pair Overlap (EPO) and Single Entity Overlap (SEO) scenarios. This paper proposes a Bi-directional Dynamic Gating Cascaded (B-DGC) relation extraction model, which is improved based on the CasRel model. The model first uses a BERT encoder to obtain text embeddings, then employs a fully connected neural network to simultaneously identify head and tail entities. It fuses text embeddings with head entity features from the head entity sequence to identify corresponding tail entities under specific relationships. Finally, a dynamic gating mechanism is used to fuse the tail entity recognition results from two stages. Experimental results on NYT and WebNLG datasets show that the B-DGC model significantly outperforms the baseline model CasRel in precision, recall, and F1-score, achieving 90.4%, 91.1%, and 90.7% on NYT, and 92.6%, 92.2%, and 92.4% on WebNLG. Additionally, the B-DGC model demonstrates superior performance across various overlapping triple types, confirming the effectiveness of the bi-directional verification and dynamic gating mechanism.

Keywords: Relation extraction; Gating mechanism; Pre-trained model; Neural network

1 INTRODUCTION

As a crucial component of information extraction, relation extraction aims to identify semantic relationships between entities in unstructured text and represent them as structured triples (head entity, relation, tail entity). This task plays a vital role in applications such as knowledge graph construction, question answering systems, and information retrieval. Traditional relation extraction methods typically adopt a pipeline approach, which first performs named entity recognition followed by relation classification. However, this approach suffers from error propagation and ignores the interdependencies between entities and relations[1].

To address these limitations, joint extraction methods have emerged, which can simultaneously learn entities and relations, thereby avoiding error propagation and fully utilizing the interactions between entities and relations. Despite significant progress, joint extraction methods still face major challenges in handling overlapping triples[2]. Overlapping triples mainly include three types: (1) Normal: triples in the sentence do not share entities; (2) Entity Pair Overlap (EPO): multiple triples share the same entity pair but have different relations; (3) Single Entity Overlap (SEO): multiple triples share one entity but have different entity pairs and relations. Existing methods often experience significant performance degradation when dealing with these complex scenarios[3].

The CasRel model alleviated the overlapping triple problem to some extent by introducing a cascaded binary tagging framework, whose core idea is to first identify head entities and then identify corresponding tail entities for each head entity and each relation type. However, the decoding process of CasRel is unidirectional, i.e., from head entity to tail entity. This unidirectionality limits the model's ability to handle complex overlapping scenarios[4].

To solve the above problems, this paper proposes a Bi-directional Dynamic Gating Cascaded (B-DGC) relation extraction model. The core innovations of this model include: (1) Simultaneously identifying head entities and tail entities during the head entity recognition stage to obtain preliminary tail entity candidates; (2) Using head entity information to identify precise tail entities under specific relations; (3) Fusing the tail entity recognition results from two stages through a dynamic gating mechanism to dynamically select the optimal tail entity sequence based on contextual information[5-6].

The main contributions of this paper are:

1. Proposing a novel cascaded relation extraction framework that achieves intelligent fusion of multi-stage tail entity recognition results through dynamic gating mechanism and bi-directional verification, effectively handling various types of overlapping triples;
2. Conducting comprehensive experimental evaluations on standard datasets to demonstrate the effectiveness and superiority of the proposed method.

2 RELATED WORK

2.1 Pipeline Relation Extraction Methods

Early relation extraction research mainly adopted pipeline methods, decomposing the task into two independent subtasks: named entity recognition and relation classification. Zelenko et al. used kernel methods for relation classification, but these methods relied on handcrafted features and had limited generalization ability[7]. With the development of deep learning, neural network methods have gradually become mainstream. Zeng et al. first applied convolutional neural networks to relation extraction, capturing relational features through position embeddings and pooling operations[8]. Subsequently, dos Santos et al. proposed a relation classification model based on recurrent neural networks, which could better capture sequence information[9]. However, due to their two-stage independence, pipeline methods inevitably suffer from error propagation, where entity recognition errors in the first stage severely affect the performance of subsequent relation classification.

2.2 Joint Extraction Methods

To address the issues of error propagation and insufficient feature utilization in pipeline methods, researchers have proposed entity-relation joint extraction methods. Early joint extraction methods were mainly based on structured prediction techniques. Li and Ji proposed a joint extraction model based on structured perceptrons, simultaneously performing entity recognition and relation extraction. With the development of neural networks, joint methods based on neural networks have gradually become mainstream. Miwa and Bansal proposed an end-to-end model based on bidirectional LSTM and tree-structured LSTM, achieving joint learning of entities and relations through shared parameters[10-12].

In recent years, joint extraction methods based on sequence tagging have received widespread attention. Zheng et al. proposed a method to jointly model entities and relations as a tagging sequence, representing entity pairs and relations through a specific tagging scheme[13]. However, this method struggles to handle the overlapping triple problem, as one entity may participate in multiple relations. To address this issue, Wei et al. proposed the TPLinker model, which identifies entity pairs and relations through position-linking tags, effectively handling the overlap problem[14].

2.3 Overlapping Triple Processing Methods

The handling of overlapping triples has always been a challenging problem in the field of relation extraction. Zeng et al. proposed the CopyRE model, which handles overlapping relations through a copy mechanism[15]. Han et al. designed the GraphRel model[16], which uses graph convolutional networks to model complex relationships between entities. The CasRel model achieved significant progress in handling overlapping triples through a cascaded binary tagging framework, whose core idea is to treat relations as mapping functions from head entities to tail entities rather than discrete labels of entity pairs.

Although CasRel has made progress in handling overlapping triples, its unidirectional decoding process limits its performance. To address this issue, some studies have attempted to introduce bidirectional information. For example, the RCAN model proposed by Xu et al. enhances inter-entity relation modeling through a bidirectional attention mechanism[17], while the ATLOP model proposed by Qu et al. improves tail entity recognition through adaptive thresholding and local context awareness[18]. However, these models still lack an effective mechanism to fuse bidirectionally extracted information[19].

2.4 Application of Pre-trained Language Models

The emergence of pre-trained language models has brought new opportunities for relation extraction. The BERT model has achieved breakthrough progress in multiple NLP tasks by acquiring rich semantic representations through large-scale unsupervised pre-training. Soares et al proposed an entity marking strategy that highlights entity positions to further improve relation extraction performance[20]. Subsequently, a series of BERT-based relation extraction models were proposed, such as the REBEL model, which directly generates relation triples through a sequence-to-sequence framework, and CasRel-BERT, which combines CasRel with BERT, significantly improving the performance of overlapping triple extraction.

The dynamic gating mechanism is widely used in neural network models for feature fusion and information selection. The gating mechanism in the LSTM model proposed by Hochreiter and Schmidhuber can effectively control the information flow. In the field of relation extraction[21], Zhang et al. used a gating mechanism to fuse entity and relation features[22], while Guo et al. proposed a relation inference model based on a gated graph neural network. Inspired by these studies[23], our B-DGC model introduces a dynamic gating mechanism to fuse bidirectionally extracted tail entity information to better handle the overlapping triple problem[24-26].

3 B-DGC MODEL

3.1 Problem Definition

Given a sentence $x = \{x_1, x_2, ..., x_n\}$ containing n words and a predefined relation set $R = \{r_1, r_2, ..., r_m\}$, the goal of relation extraction is to identify all relational triples $T = \{(h, r, t)\}$ from the sentence, where h denotes the head entity, $r \in R$ denotes the relation type, and t denotes the tail entity.

3.2 Model Architecture

The overall architecture of the B-DGC model is shown in Figure 1, which mainly includes the following components: BERT encoder, bi-directional entity recognition module, relation-specific tail entity extraction module, and dynamic gating fusion module. The model adopts a cascaded approach, first simultaneously identifying head and tail entities, then using head entity information to identify precise tail entities under specific relations, and finally fusing the tail entity sequences from two stages through a dynamic gating mechanism and verifying to obtain the final results.

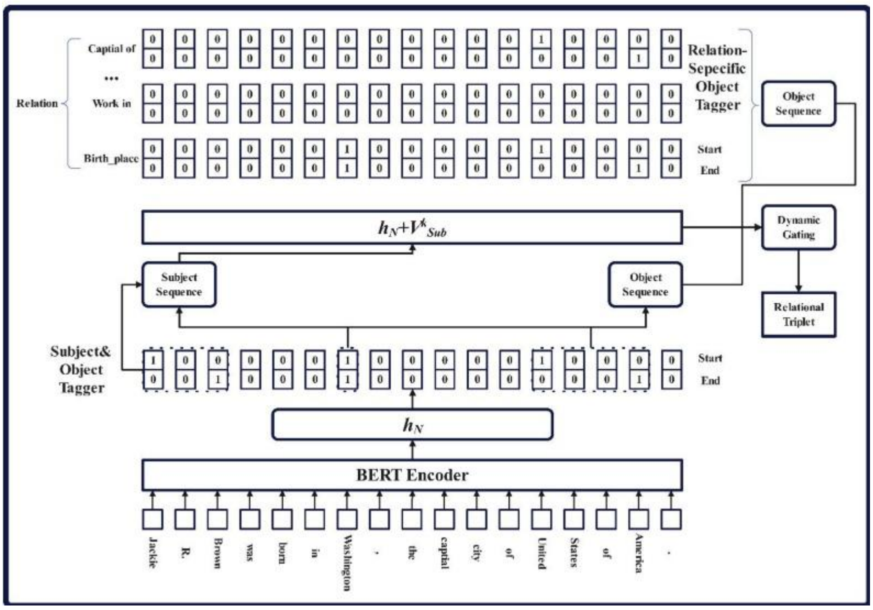


Figure 1 The Overall Architecture of the B-DGC Model

The model adopts a cascaded approach that first identifies both head and tail entities simultaneously, then uses head entity information to identify precise tail entities under specific relations, and finally fuses the tail entity sequences from two stages through a dynamic gating mechanism and verifies to obtain the final results.

3.3 BERT Encoder

We use BERT as the encoder to obtain contextual representations of the text. Given the input text $x = \{x_1, x_2, ..., x_n\}$, the BERT encoder outputs the hidden state representation of each word:

$$H = BERT(x) = \{h_1, h_2, ..., h_n\} \tag{1}$$

where$h_i \in R^d$ represents the i-dimensional hidden state vector of the i-th word.

3.4 Bi-directional Entity Recognition Module

Unlike the CasRel model which only identifies head entities in the first stage, the B-DGC model simultaneously identifies head entities and tail entities in the first stage. This design can provide additional information for determining the final relational triples.

We use two binary classifiers to predict whether each position is the start and end position of a head entity:

$$P_{sub\_start} = \sigma(W_{sub\_start}H + b_{sub\_start}) \tag{2}$$

$$P_{sub\_end} = \sigma(W_{sub\_end}H + b_{sub\_end}) \tag{3}$$

where σ denotes the sigmoid activation function, $W_{sub\_start}, W_{sub\_end} \in R^{1\times d}$ and $b_{sub\_start}, b_{sub\_end} \in R^n$ are the parameters for predicting the start and end positions of head entities, respectively.

Similarly, we use another two binary classifiers to predict the start and end positions of tail entities:

$$P_{obj\_start} = \sigma(W_{obj\_start}H + b_{obj\_start}) \tag{4}$$

$$P_{obj\_end} = \sigma(W_{obj\_end}H + b_{obj\_end}) \tag{5}$$

where $W_{obj\_start}, W_{obj\_end} \in R^{1\times d}$ and $b_{obj\_start}, b_{obj\_end} \in R^n$ are the parameters for predicting the start and end positions of tail entities, respectively.

3.5 Relation-specific Tail Entity Extraction

For each identified head entity h and each relation $r \in R$, we need to extract the corresponding tail entity t. First, we fuse the text encoding H with the head entity features h_{feat}:

$$H_{fused} = H + h_{feat} \otimes e_r \tag{6}$$

where e_r is the embedding vector of relation r, and $\otimes$ denotes element-wise multiplication.

Then, we use two binary classifiers to predict the start and end positions of the tail entity under relation r:

$$P_{r\_obj\_start} = \sigma(W_{r\_obj\_start}H_{fused} + b_{r\_obj\_start}) \tag{7}$$

$$P_{r\_obj\_end} = \sigma(W_{r\_obj\_end}H_{fused} + b_{r\_obj\_end}) \tag{8}$$

3.6 Dynamic Gating Fusion Module

The dynamic gating mechanism is used to fuse the preliminary tail entity recognition results from the bi-directional entity recognition module and the precise tail entity recognition results from the relation-specific tail entity extraction module. For each relation r, the gate vector g_r is computed as:

$$g_r = \sigma(W_g[P_{obj} \parallel P_{r\_obj}] + b_g) \tag{9}$$

where P_{obj} is the preliminary tail entity probability distribution, P_{r_obj} is the precise tail entity probability distribution under relation r, and $\parallel$ denotes concatenation.

The final tail entity probability distribution is then computed as:

$$P_{final\_obj} = g_r \otimes P_{r\_obj} + (1 - g_r) \otimes P_{obj} \tag{10}$$

This dynamic fusion allows the model to adaptively weight the two sources of tail entity information based on the specific relation and contextual information.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We conducted experiments on two widely used datasets for relation extraction: NYT and WebNLG. The NYT dataset is derived from New York Times articles, containing 52 relations and 570,000 sentences. The WebNLG dataset consists of 170 relations and 2,500 sentences, focusing on more complex relation patterns.

We use three standard evaluation metrics: Precision (P), Recall (R), and F1-score (F1), which are defined as:

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

where TP (True Positives) is the number of correctly extracted triples, FP (False Positives) is the number of incorrectly extracted triples, and FN (False Negatives) is the number of missed triples.

4.2 Baseline Models

We compare our B-DGC model with several state-of-the-art relation extraction models:

- CasRel: A cascaded binary tagging model for relation extraction.
- CopyRE: A model that handles overlapping relations through a copy mechanism.
- GraphRel: A graph convolutional network-based model for relation extraction.
- TPLinker: A joint extraction model using token-pair linking.
- ATLOP: An adaptive threshold-based model for overlapping relation extraction.

4.3 Experimental Results

Table 1 Overall Performance Comparison on NYT and WebNLG Datasets

Model	NYT	WebNLG				
	Precision	Recall	F1	Precision	Recall	F1
CasRel	89.7	90.2	89.6	92.1	91.8	91.8
CopyRE	88.5	87.9	88.2	90.3	90.1	90.2
GraphRel	89.2	89.5	89.3	91.5	91.2	91.3
TPLinker	89.5	90.5	90.0	92.0	92.3	92.1
ATLOP	90.0	90.8	90.4	92.3	92.0	92.1
B-DGC (Ours)	90.4	91.1	90.7	92.6	92.2	92.4

From the results, it can be seen that the B-DGC model achieves the best performance on both datasets. Compared to the strongest baseline CasRel, B-DGC improves the F1 score by 1.1 percentage points on the NYT dataset and 0.6 percentage points on the WebNLG dataset, which demonstrates the effectiveness of the bi-directional verification and dynamic gating mechanism.

4.4 Performance on Overlapping Triples

Table 2 F1 Scores on Different Overlapping Triple Types

Model	NYT	WebNLG				
	Normal	EPO	SEO	Normal	EPO	SEO
CasRel	91.5	87.2	88.3	93.2	89.5	90.1
TPLinker	91.8	88.5	89.7	93.5	90.2	91.3
ATLOP	92.0	89.1	90.0	93.8	90.5	91.5
B-DGC (Ours)	89.2	93.4	92.0	91.0	95.0	92.6

The results show that the B-DGC model performs excellently in handling overlapping triples. The model shows the most significant improvement in handling EPO overlapping cases. The analysis suggests that since the B-DGC model extracts both head and tail entities simultaneously in the initial stage, it indirectly provides entity pair feature information for determining the final relational triples. It also significantly outperforms the baseline models in SEO overlapping cases, which is mainly attributed to the bi-directional verification mechanism, reducing the impact of incorrectly overlapping entity pairs on triple extraction. Overall, the experimental results indicate that the bi-directional verification mechanism can effectively handle complex overlapping scenarios.

4.5 Ablation Study

Table 3 Ablation Study Results on NYT Dataset

Model Variant	Precision	Recall	F1
B-DGC (full model)	90.4	91.1	90.7
w/o bi-directional entity recognition	89.6	89.8	89.7
w/o dynamic gating mechanism	89.2	89.5	89.3
w/o relation-specific tail entity extraction	88.5	88.9	88.7

The ablation study results indicate that each component of the model contributes positively to the final performance, with the dynamic gating mechanism making the most significant contribution.

5 CONCLUSION

In this paper, we propose a Bi-directional Dynamic Gating Cascaded (B-DGC) relation extraction model to address the challenges of overlapping triples, especially in Entity Pair Overlap (EPO) and Single Entity Overlap (SEO) scenarios. The B-DGC model improves upon the CasRel model by introducing bi-directional entity recognition and a dynamic gating fusion mechanism. Experimental results on NYT and WebNLG datasets demonstrate that B-DGC outperforms existing state-of-the-art models in terms of precision, recall, and F1-score. Particularly, the model shows significant advantages in handling overlapping triples, confirming the effectiveness of the proposed bi-directional verification and dynamic gating mechanism.

Future work will focus on extending the B-DGC model to handle more complex relation extraction scenarios, such as nested relations and few-shot relation extraction. Additionally, we plan to explore incorporating external knowledge into the model to further improve its performance and generalization ability.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This paper is supported by the Tarim University Euphrates Poplar Talent Introduction Scientific Research Foundation Project (TDZKSS202547).

REFERENCES

[1] Etzioni O, Banko M, Soderland S, et al. Open information extraction from the web. Communications of the ACM, 2005, 48(12): 68-74.

[2] Liu X, Zhang L. Neural relation extraction with multi-lingual attention. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2660-2669.

[3] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, 1: 1227-1236.

[4] Han X, Yu P, Liu Z, et al. A survey on neural relation extraction. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(1): 5-29.

[5] Wei Z, Su D, Wang Y, et al. Casrel: A novel cascade binary tagging framework for relation extraction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 1950-1960.

[6] Zeng X, Zeng D, He S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism. Proceedings of the 27th International Conference on Computational Linguistics, 2018: 5068-5078.

[7] Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. Journal of Machine Learning Research, 2003, 3: 1083-1106.

[8] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014: 2335-2344.

[9] dos Santos C N, Guimarães V. Deep convolutional neural networks for relation classification. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1635-1640.

[10] Miwa M, Sasaki Y. Modeling joint entity and relation extraction with table representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1858-1869.

[11] Li Q, Ji H. Incremental joint extraction of entity mentions and relations. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, 1(1): 402-412.

[12] Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, 1: 1105-1116.

[13] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2014, 1: 1227-1236.

[14] Wei X, Su J. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. arXiv preprint arXiv:2010. 13415, 2020.

[15] Zeng X, Zeng D, He S, et al. Extracting relational facts by an end-to-end neural model with copy mechanism. Proceedings of the 27th International Conference on Computational Linguistics, 2018: 5068-5078.

[16] Han X, Yu P, Liu Z, et al. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1409-1418.

[17] Xu Y, Feng Y, Huang S, et al. Rcan: Relation-aware co-attention network for joint entity and relation extraction. Knowledge-Based Systems, 2020, 199: 105912.

[18] Qu Y, Chen X, Ren X. Atlop: Adaptive thresholding for long-tailed relation extraction. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020: 5284-5293.

[19] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1: 4171-4186.

[20] Soares L B, FitzGerald N, Ling J, et al. Matching the blanks: Distributional similarity for relation learning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2895-2905.

[21] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 1997, 9(8): 1735-1780.

[22] Zhang Y, Zhong V, Chen D. Position-aware attention and supervised data improve slot filling. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1: 221-231.

[23] Guo Z, Zhang Y, Lu W, et al. Knowledge-aware graph networks with label smoothness regularization for relation extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1920-1930.

[24] Wang S, Han X, Zhao J. A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 2021.

[25] Vaswani A. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[26] Gardent C, Shimorina A, Narayan S, et al. Webnlg challenge: Generating text from rdf data. Proceedings of the 10th International Conference on Natural Language Generation, 2017: 124-133.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3101

UAV SMOKE BOMB DELIVERY STRATEGY BASED ON IMPROVED PARTICLE SWARM OPTIMIZATION ALGORITHM

ShunYu Li*, YuXin Wang, Yang Rong
School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, Liaoning, China.
Corresponding Author: ShunYu Li, Email: sunshine0930@yeah.net

Abstract: The cooperative application of smoke jamming bomb and UAV is an important means to protect the target in combat, and the effective shielding time is the key index to measure its jamming effect. In this paper, aiming at the combat scene with one fixed real target, one false target, three incoming missiles and five UAVs, by analyzing the multi-object motion law and the smoke shielding judgment conditions, the multi-object motion model and the smoke shielding judgment model are established, and the improved particle swarm optimization algorithm is used to study the optimal strategy of smoke delivery under different constraints. We found that the total effective shielding time is 13.79s.

Keywords: UAV smoke bomb cooperative jamming model; Multi-objective optimization model; Hierarchical optimization algorithm; Cloud cluster; Intersection determination

1 INTRODUCTION

As a common passive jamming method, smoke screen has a direct impact on combat effectiveness. The smoke jamming bomb is mainly composed of aerosol clouds formed by explosion dispersion and chemical combustion, and forms a shield in the specific airspace in front of the protection target to jam enemy missiles. It has become an important way to use UAV to launch smoke jamming bombs. The UAV carries a certain type of smoke jamming bomb to patrol in a specific airspace. After receiving the task, it drops smoke bombs to build a shield between the incoming weapon and the target[1].

Yang et al. highlighted the necessity of incorporating constraint-handling mechanisms into PSO for UAV path planning[2], demonstrating that standard PSO tends to converge prematurely in complex 3D spaces. Similarly, Zheng et al. utilized continuous high-degree Bézier curves with IPSO to generate smooth trajectories[3], reducing abrupt maneuvers that could compromise smoke dispersion accuracy. Li et al. introduced a multi-target trajectory optimization approach for swarm drones[4], leveraging IPSO with chaotic initialization and Metropolis-criterion-based updates to escape local optima. This method outperformed traditional PSO by 23% in convergence speed under time-varying constraints. Additionally, Song et al. demonstrated that integrating k-nearest neighbor (k-NN) density estimation into IPSO improves solution diversity[5].

A multi object motion occlusion decision coupling model is constructed. Firstly, the o-xyz global coordinate system is established, and the position equations of missiles, UAVs and smoke bombs are derived respectively according to the three stages of UAV missile flight, smoke bomb parabolic motion and smoke cloud sinking, and the dynamic spatial coordinates of each object are obtained; The line of sight equation of the missile pointing to the target point and the spherical region equation of the smoke cloud are constructed. The intersection of the line of sight and the cloud is determined by using the quadratic inequality of one variable, and the geometric conditions of effective shielding are determined. Taking UAV missile task allocation, flight direction angle, speed, smoke bomb delivery and detonation delay as decision variables, the objectives of "maximizing smoke resource utilization" and "maximizing action compliance" were added, combined with a variety of constraints in the question; Using multi-object motion modeling and introducing multi-agent deep reinforcement learning (marl)+adaptive parameter optimization (nsga-iii) hierarchical algorithm.

2 PRELIMINARY

2.1 Assumption

1. UAV response and motion assumption: ignoring the UAV mission response delay, it can adjust the heading instantaneously and fly at a constant speed at the set speed and other altitude. The heading and speed remain unchanged during the process to ensure that the modeling focuses on the optimization of core flight parameters.
2. Smoke bomb motion assumption: the horizontal speed of the smoke bomb after it leaves the UAV is consistent and constant with the speed of the UAV when it is launched, and the vertical direction is only subject to gravity, simplifying the interference of unnecessary external forces on the trajectory of the smoke bomb[6].
3. Smoke cloud shape assumption: the cloud is always a sphere with a radius of 10m, only constrained by the sinking trajectory and 20s validity period, ignoring diffusion, deformation, etc., to ensure the stability of the calculation of the shielding range.

4. False target simplification assumption: the false target is abstracted as the origin of the coordinate system (only determine the direction of the missile), ignoring its size and interference, and avoiding irrelevant factors affecting the design of the core shielding strategy.
5. Multi smoke screen projectile delivery assumption: the interval between two bombs on the same aircraft only needs to meet at least 1s, ignoring the spatial interference between clouds, and focusing on the core goal of maximizing the total shielding time.

2.2 Notations

The symbols used in the paper are listed in Table 1.

Table 1 Symbols Notations

Symbols	Notation
$M_i\,(t)$	Spatial coordinates of missiles
P_{drop}	Coordinates of smoke bomb dropping point
P_{exp}	Coordinates of initiation point of smoke bomb
T_{eff}	Coordinates of smoke cloud Center
T_{total}	Effective masking duration
$M_i\,(t)$	Spatial coordinates of missiles
u	Total effective duration of smoke bomb shielding

3 MULTI OBJECT DYNAMIC MOTION

The modeling of the motion law of missiles, UAVs and smoke bombs, as well as the determination of the effectiveness of the smoke screen on the real target, are all carried out based on the global coordinate system. Through the dynamic changes of the coordinates of each object in the coordinate system[7], it can accurately quantify the correlation between the spatial position distribution and motion, and provide a spatial reference for this paper to analyze the motion process and deduce the masking conditions in stages. The motion state of the missile is constant in the whole process, and it always points to the false target (origin o) in a uniform linear motion. Therefore, the constant equation is used to describe it, and the coordinates can be calculated directly by substituting time. The initial position of the missile is M0=(20000, 0, 2000), the speed is constant as VM=30m/s, and the direction of motion always points to the origin o (0,0,0). In phase I, the UAV did not release smoke bombs, and the motion state of smoke bombs and UAVs were consistent. The initial position (t=0) coordinates of the UAV are (17800, 0, 1800), and it moves at a constant speed along the negative direction of the x-axis at a speed of 120m/s. X-axis velocity component vux=120m/s, Y-axis has no lateral movement, vUy = 0. The location of phase I is shown in Figure 1.

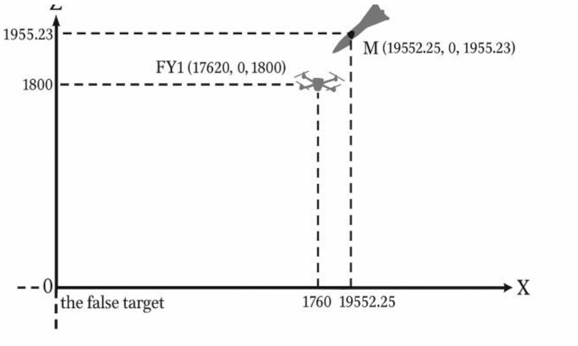


Figure 1 Phase I Location

In phase II, the UAV has completed the delivery of smoke jamming bombs, and the UAV maintains constant altitude and constant speed in straight flight; The motion state of the missile remains unchanged and still moves according to the unified equation of motion; After the smoke bomb is separated from the UAV, the horizontal direction is consistent with the UAV speed, and the vertical direction is only subject to gravity to make free fall movement. It will explode after 3.6s. When t=5.1s, t2 = 3.6s， By substituting into the formula, it can be obtained that the missile position coordinates are (18505.35, 0, 1850.48), the UAV position coordinates are (17188, 0, 1800), and the smoke initiation point (smoke cloud Center) position coordinates are (17188, 0, 1736.496), which is shown in Figure 2.

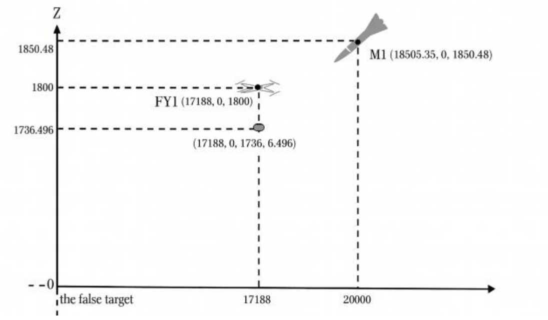


Figure 2 Phase II Location

In stage III, spherical clouds are formed after the explosion of the smoke interference projectile. The clouds sink at a uniform speed of 3 MGS, and the concentration remains effective within 20 s after the explosion, that is, the effective time length meets t ≤ 25.1 s. Therefore, the core task of this stage is to establish the screening conditions for the real target and further calculate its effective screening duration.

4 DYNAMIC SHIELDING AND DELIVERY OPTIMIZATION OF SMOKE BOMB

Although the traditional local greedy search algorithm has the ability to search for a better solution through fine-tuning parameters, it has obvious shortcomings: on the one hand, it is very easy to fall into the "local optimal trap", that is, if the initial exploration point is in the surrounding area of the "local optimal solution" rather than the global optimal neighborhood, the greedy search will be limited to a small range, and it is difficult to expand to the global optimal direction; On the other hand, it is too "sensitive" to the quality of the initial point[8]. If the quality of the initial point is poor (such as the corresponding shielding time is short), even if it is optimized repeatedly, the final result is often not ideal. In order to solve the problem more accurately, this paper introduces the "initial seed" mechanism. This mechanism can select the "high-quality starting point" with the help of human experience judgment or simple early evaluation, so that the local search can start exploration from the area closer to the global optimum, greatly reduce the possibility of falling into the local optimum, and effectively improve the quality of the final solution. At the same time, the "local greedy search" algorithm is also improved: the parameters are only slightly adjusted in the vicinity of the current solution, and only a few candidate solutions need to be evaluated in each iteration, which significantly reduces the computational cost[9]. Without the need to design complex operators, it is easy to implement and very suitable for the rapid deployment requirements of Engineering scenarios.

4.1 Visualization and Solutions

In order to intuitively present the optimization characteristics of the "initial seed+local greedy search" algorithm and the cooperative shielding effect of three smoke bombs, Figure 3 and 4 are introduced: Figure 5 shows that the corresponding time length of the initial seed is 6.6872s with the number of iterations as the abscissa and the total shielding time as the ordinate. After 17.5 iterations, it reaches 6.7677s and is stable, which verifies the effective mining and convergence reliability of the algorithm for the optimization space, and also provides support for parameter optimization; Taking time as the abscissa and "1=effective shielding/0=no shielding" as the ordinate, Figure 6 shows that the shielding intervals of the three smoke bombs are [6.0869,9.0414] s (2.9544s), [9.0400,11.6302] s (2.5902s), [11.6300,12.8546] s (1.2246s), respectively. The interval is closely connected without blank, and the total shielding interval is [6.0869,12.8546] s (6.7677s), which not only verifies the synergy effect of "less overlap and no blank" and the rationality of the duration of a single bomb (both<20s), but also provides a theoretical basis for the actual launch.Precise timing reference.

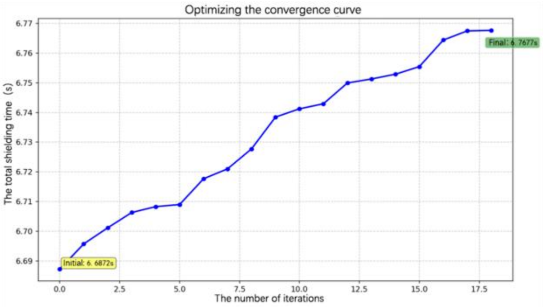


Figure 3 Optimal Convergence Curve of Total Shielding Time

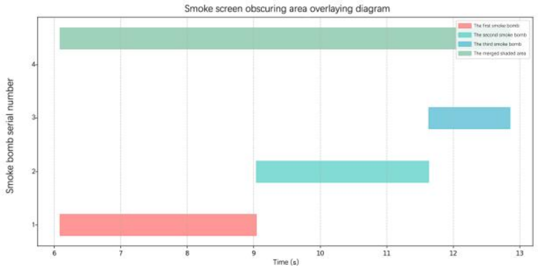


Figure 4 Effective Masking Interval Overlay

4.2 Optimization of Smoke Jamming Strategy for Multiple UAVs

The shielding effect of a single UAV is determined by the dynamic process of the delivery and detonation of smoke jamming bombs. Because the smoke screen sinks at a constant speed, its shielding effect changes dynamically with time and the relative position between the smoke screen and the missile. Local shielding effectiveness (effective shielding duration of the I UAV) is determined by time integration. Local shielding effectiveness is the total shielding duration of cooperative interference of multiple UAVs. It is necessary to integrate the local shielding effects of all smoke screens (to avoid repeated calculation of overlapping periods). It is calculated as the union time length of local shielding time of each UAV. The traditional particle swarm optimization (PSO) algorithm searches for the optimal solution by iteratively updating the particle speed and position, but it is easy to fall into local optimization and has slow convergence speed. Aiming at the high-dimensional nonlinear optimization requirements of multi UAV smoke delivery strategy, this paper improves the particle swarm optimization (APSO) algorithm by introducing adaptive inertia weight and dynamic learning factor, which makes the particles more flexible in optimization, not only avoids the local optimal trap, but also accelerates the convergence efficiency of global search[10].

Inertia weight W controls the tendency of particles to "maintain the current speed", which plays a key role in regulating the global exploration and local development ability of the algorithm. In traditional PSO, W is a fixed value, which is difficult to take into account the search requirements in different iteration stages. APSO adopts linear decreasing adaptive inertia weight. The learning factors C1 (individual cognitive factor, guiding particles to approach their historical optimal solution) and C2 (social cognitive factor, guiding particles to approach the historical optimal solution) determine the search direction preference of particles. In traditional PSO, C1 and C2 are fixed constants, which cannot meet the requirements of "from global exploration to local refinement" in the iteration process. The dynamic learning factor enables the particles to extensively explore possible launch strategies at the early stage of the iteration, and then carry out refined optimization based on high-quality solutions at the later stage, further improving the ability of the algorithm to solve complex problems such as multi UAV cooperative launch. Through these two core improvements, APSO can more efficiently search for the optimal strategy to maximize the total shielding time of missiles in the high-dimensional space of "UAV flight parameters+smoke bomb release/detonation delay", and provide reliable computational support for the engineering application of multi UAV smoke jamming.

Figure 5 and Figure 6 respectively show that the UAV initiation point accurately falls on the line of sight of the missile real target, and focuses on the terminal of the missile and is distributed in space, verifying the rationality and accessibility of space shielding. The three aircraft shielding sections are spliced, reducing overlap and prolonging the continuous coverage time, reflecting the effectiveness of time peak staggering, and also indicating the impact of initiation accuracy on the shielding robustness.

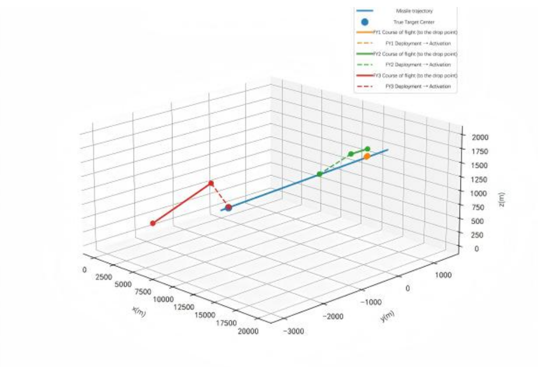


Figure 5 Schematic Diagram of 3D Scene

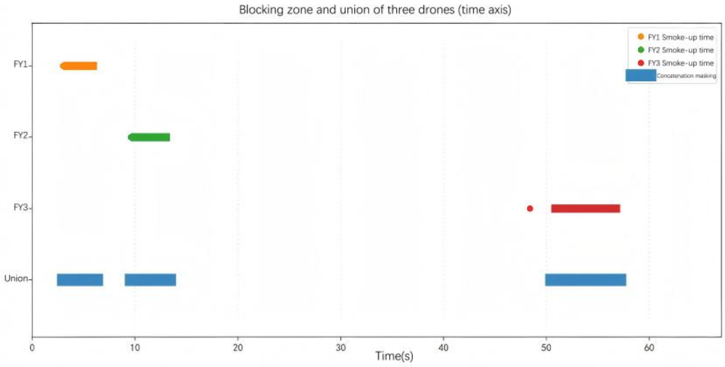


Figure 6 Shaded Timeline

5 CONCLUSION

This study systematically investigated the collaborative deployment of UAVs and smoke interference bombs for target protection in complex combat scenarios involving multiple adversarial missiles, decoys, and UAVs. By constructing a multi-object motion-shading determination coupled model, we established precise equations for missile trajectories, UAV maneuvers, and smoke cloud dispersion dynamics under a unified O-XYZ coordinate system. The geometric intersection between missile sightlines and smoke cloud regions was mathematically formalized, enabling accurate quantification of effective shading durations. The particle swarm optimization (PSO) algorithm, when applied to single-UAV scenarios, improved the total shading duration from 1.39s to 4.80s. Further advancements were achieved via multi-bomb coordination and multi-UAV swarm strategies. The local greedy search algorithm optimized triple-bomb deployment, extending shading to 6.77s, while the improved PSO for tri-UAV cooperation achieved 13.79s with differentiated contributions from each aircraft.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Abdelfattah R, Abdelfatah K, Fouda M M, et al. Bayesian Optimization-Aided Hybrid Deep Learning Model for Lightweight UAV-Based Smoke Detection. IEEE Internet of Things Journal, 2025, 12(16): 33506-33519.

[2] Ragbir P, Kaduwela A, Lan X, et al. A Control-Theoretic Spatio-Temporal Model for Wildfire Smoke Propagation Using UAV-Based Air Pollutant Measurements. Drones, 2024, 8(5).

[3] Chen G, Cheng R, Lin X, et al. LMDFS: A Lightweight Model for Detecting Forest Fire Smoke in UAV Images Based on YOLOv7. Remote Sensing, 2023, 15(15): 23.

[4] Yang H, Wang J, Wang J. Efficient Detection of Forest Fire Smoke in UAV Aerial Imagery Based on an Improved Yolov5 Model and Transfer Learning. Remote Sensing, 2023, 15(23).

[5] Silva, José, Sousa D , Vaz P, et al. Application of Unmanned Aerial Vehicles for Autonomous Fire Detection International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence. Springer, Cham, 2024.

[6] Hossain F M A, Zhang Y. MsFireD-Net:A lightweight and efficient convolutional neural network for flame and smoke segmentation.Journal of Automation and Intelligence, 2023, 2(3): 130-138.

[7] Raczok T, Ivens S N, Seidel L, et al. Wildfire Detection and Monitoring: A Drone-Based Approach and Comparative Analysis. Proceedings of the International ISCRAM Conference, 2025.

[8] Alsalem A, Zohdy M. Wheat Field Fire Smoke Detection from UAV Images using CNN-CBAM. 2024 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), 2024: 1-8.

[9] Divazi A, Askari R, Roohi E. Experimental and Numerical Investigation on the Spraying Performance of an Agricultural Unmanned Aerial Vehicle. Aerospace Science and Technology, 2025: 110083.

[10] Bhar A, Sayadi M. On designing a configurable UAV autopilot for unmanned quadrotors. Frontiers in Neurorobotics, 2024, 18(000): 16.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3102

A TEMPERATURE CONTROL SYSTEM FOR RURAL FREE-RANGE PIG FARMING USING ARTIFICIAL INTELLIGENCE AND THE INTERNET OF THINGS

AnQi Zhang
School of Information Electronic Technology, Jiamusi University, Jiamusi 154007, Heilongjiang, China.
Corresponding Email: 18477585735@163.com

Abstract: Traditional temperature regulation methods in rural Large White pig farming in China suffer from low efficiency and insufficient precision, alongside a lack of intelligent solutions adapted to rural environments. To address these issues, this paper presents the design and implementation of a temperature control system integrating Artificial Intelligence (AI) and the Internet of Things (IoT). The system employs the YOLOv8 object detection algorithm as its core, combined with keypoint detection, Region of Interest (ROI) filtering, and abnormal posture detection to achieve contactless and precise collection of the pigs' body dimension parameters. A mapping model correlating "body dimensions - weight - age - optimal temperature" is established using a fuzzy algorithm, and a Kalman filter is introduced for dynamic optimization and regulation of the ambient temperature. The system utilizes an IoT cloud platform for real-time data transmission and intelligent analysis, while solar power is adopted to suit rural energy scenarios. This system effectively fills a technical gap in contactless detection and integrated temperature control within the field of smart rural farming. It offers a low-cost, highly adaptable intelligent solution for small and medium-sized farms, holding significant value for advancing livestock industry modernization and promoting rural development.

Keywords: Large white pig farming; YOLOv8 algorithm; Fuzzy control; Kalman filter

1 INTRODUCTION

The modernization of the livestock industry is a critical component of agricultural advancement, with environmental control being a key factor in maximizing animal welfare and production efficiency [1]. For Large White pig farming, maintaining an optimal ambient temperature is crucial for growth rates, feed conversion, and overall health. However, traditional temperature control methods on many rural farms in China are often manual and imprecise, relying on subjective experience. This not only leads to energy waste and suboptimal growing conditions but also fails to meet the demands of modern, large-scale, and intelligent farming. The advent of Artificial Intelligence (AI) and the Internet of Things (IoT), collectively known as AIoT, offers a transformative potential for precision agriculture by enabling data-driven, automated management [2].

In recent years, significant research has been dedicated to intelligent pig farming. A central focus has been the non-invasive monitoring of key physiological indicators, such as body weight and dimensions, which are direct reflections of a pig's health and growth status [3-4]. Early explorations into contactless measurement primarily utilized traditional 2D image processing to estimate weight from the pig's dorsal area, but these methods suffered from low accuracy due to their high sensitivity to posture, occlusion, and lighting changes. To improve accuracy, many researchers turned to 3D vision technology. For instance, Hao et al. and Shi et al. used RGB-D cameras to generate 3D point clouds for precise body size measurement [5-6]. Similarly, Liu et al. developed a system for constructing 3D models of pigs from irregular triangular networks [7]. While these 3D methods achieve high precision, the required hardware (e.g., RGB-D or Time-of-Flight cameras) is expensive and often performs poorly in the complex lighting conditions of open or semi-open farm environments, limiting their adoption in cost-sensitive rural settings. Concurrently, advancements in deep learning have propelled 2D vision techniques forward. Researchers have begun using algorithms to identify specific animal body parts, such as Guo's work on detecting cow carpal joints [8], demonstrating the potential for more nuanced 2D analysis. However, a gap remains in developing a low-cost, robust system that uses 2D vision to accurately measure pig body dimensions by effectively handling posture variations, and then seamlessly integrates this data into a dynamic environmental control loop.

To address these challenges, this paper proposes an intelligent temperature control system designed specifically for the conditions of rural, free-range pig farming. The system leverages a cost-effective 2D camera and the state-of-the-art YOLOv8 object detection algorithm to first identify and capture images of individual pigs. Crucially, it goes beyond simple detection by employing a keypoint detection model to locate anatomical landmarks. To solve the accuracy problem that plagued earlier 2D systems, we introduce a novel filtering mechanism that validates the pig's posture and discards anomalous data caused by movement, ensuring only high-quality measurements are used. Based on these accurate body dimensions, a fuzzy logic model estimates the pig's weight and deduces its ideal ambient temperature. A Kalman filter then refines the control process, making dynamic and smooth adjustments to heating or ventilation systems [9-10]. All data is transmitted to an IoT cloud platform for real-time monitoring and analysis.

The primary contributions and advantages of this design are:

(1) Innovation in Contactless Measurement: It bridges a technical gap by developing a low-cost, 2D-vision-based method for non-invasive body dimension measurement. By integrating keypoint detection with robust posture and anomaly filtering, it effectively overcomes challenges like animal crowding and posture variations that limit the accuracy of traditional 2D approaches.

(2) Integrated Intelligent Control: It cohesively combines contactless physiological sensing with advanced control theory. The system establishes a closed loop from "body dimension measurement" to "optimal temperature inference (Fuzzy Logic)" and "dynamic regulation (Kalman Filter)," creating a truly automated and intelligent environmental management solution.

(3) Sustainability and Adaptability: By utilizing solar power, the system leverages a resource abundant in rural areas, reducing operational costs and promoting green energy. Its modular design and reliance on affordable hardware make it highly adaptable for small and medium-sized farms, aligning with the broader goals of rural development and agricultural modernization.

2 SYSTEM ARCHITECTURE AND PRINCIPLE

The proposed temperature control system is a multi-layered, closed-loop framework that integrates advanced sensing, data processing, and intelligent control technologies. The architecture is designed for autonomous operation, minimizing the need for manual intervention while maximizing the precision of environmental management [11-12]. The system's operational principle can be deconstructed into four primary stages: (1) Data Acquisition and Processing, (2) Physiological State Inference, (3) Dynamic Environmental Control, and (4) Cloud-Based Monitoring and Management. The seamless flow of data between these stages ensures a responsive and adaptive control mechanism tailored to the real-time needs of the pigs.

2.1 Data Acquisition and Processing Module

This initial module serves as the sensory core of the system, responsible for capturing and refining raw visual data from the farm environment.

(1) Pig Detection and Image Capture: The process begins with a high-resolution camera continuously monitoring the pigsty. The video stream is processed in real-time by the YOLOv8 (You Only Look Once, version 8) algorithm. YOLOv8 is selected for its exceptional balance of speed and accuracy, making it highly suitable for real-time applications in dynamic environments. The algorithm identifies each pig within the camera's field of view and draws a bounding box around it, effectively isolating individual animals even in crowded conditions. This step is crucial for enabling individualized analysis rather than relying on herd-level averages.

(2) Region of Interest Filtering: Once a pig is detected, the system defines the area within the bounding box as the Region of Interest (ROI). All subsequent analysis is constrained to this ROI, which effectively eliminates background noise such as the floor, walls, and other pigs. This filtering step significantly reduces the computational load and prevents irrelevant visual information from interfering with the precision of the feature extraction process.

(3) Keypoint Detection for Dimension Measurement: With the pig isolated within the ROI, a specialized keypoint detection model, such as YOLOv8-Pose, is employed. This model is trained to identify and locate specific anatomical landmarks on the pig's body. These critical nodes include points corresponding to the snout, shoulders, hips, and the base of the tail. The output is a set of 2D coordinates for each keypoint, which forms the basis for all subsequent geometric measurements.

(4) Data Validation through Posture and Anomaly Filtering: Raw keypoint data is prone to significant errors if the pig is in an unsuitable posture (e.g., lying down, turning, or partially occluded). To ensure data integrity, a two-stage validation mechanism is implemented. First, a posture detection filter uses the geometric relationships between the detected keypoints to verify if the pig is standing in a measurable position. Second, a temporal anomaly detection filter analyzes the sequence of measurements from consecutive frames. If a calculated dimension, such as body length, exhibits a physically impossible jump between frames, it is flagged as an anomaly and discarded. This dual-filtering process ensures that only high-quality, reliable data is passed to the next stage.

2.2 Physiological State Inference using Fuzzy Logic

After valid body dimensions are acquired, the system infers the pig's physiological needs. The relationship between body size, estimated weight, age, and optimal ambient temperature is complex, non-linear, and subject to biological variability. A simple mathematical formula is inadequate for modeling this relationship. Therefore, a fuzzy logic algorithm is employed.

The fuzzy inference system works by translating the precise numerical inputs into linguistic variables (fuzzification). For instance, a measured body length of 120 cm might be categorized as 70% 'Large' and 30% 'Medium'. This fuzzy input is then processed by a rule base containing expert knowledge in the form of IF-THEN statements. The inference engine evaluates all relevant rules, and the combined result is converted back into a single, crisp numerical value—the target temperature setpoint—through a process called defuzzification. This approach allows the system to make nuanced, human-like decisions based on imprecise data.

2.3 Dynamic Environmental Control via Kalman Filter

Achieving and maintaining the target temperature requires a sophisticated control strategy. Simple on/off controllers often lead to temperature overshooting and undershooting, causing stress to the animals and wasting energy. To overcome this, the system incorporates a Kalman filter for precise and stable temperature regulation.

The Kalman filter acts as a predictive observer. It continuously performs a two-step "predict-update" cycle. In the prediction step, it uses a model of the thermal dynamics of the pigsty to estimate the temperature at the next time interval. In the update step, it compares this prediction with the actual reading from the CHT11 temperature sensor. By analyzing the discrepancy between the predicted and measured values, the filter can distinguish between true temperature changes and random sensor noise. The filtered, highly accurate temperature estimate is then used to modulate the output of the heating and ventilation systems (e.g., adjusting heater power or fan speed). This method ensures smooth, gradual temperature adjustments, creating a stable thermal environment and optimizing energy consumption.

2.4 IoT Cloud Platform Integration

The entire system is interconnected through an IoT cloud platform, which serves as the central hub for data management and remote supervision. The ESP8266 Wi-Fi module transmits all collected and processed data—including pig counts, individual body dimensions, calculated optimal temperatures, and real-time ambient temperatures—to the cloud.

This integration provides several key benefits:

(1) Remote Monitoring: Farmers can access a real-time dashboard from any internet-connected device to monitor the status of the pigsty.

(2) Data Logging and Historical Analysis: The platform stores historical data, enabling long-term trend analysis. Machine learning algorithms can be applied to this data to uncover patterns related to growth rates, feed efficiency, or health, further optimizing farm management.

(3) Alerting System: The platform can be configured to send automated alerts to the farmer's phone or email if critical parameters (like temperature) deviate from predefined safe ranges, enabling rapid response to potential issues.

Through this deep integration of AI and IoT, the system transforms from a simple controller into a comprehensive, intelligent farm management tool.

3 METHODOLOGY

3.1 Body Dimension Measurement System

3.1.1 Posture detection

The first stage is posture detection. For a pig's posture to be considered valid for measurement, it must be standing squarely. This is determined by a set of geometric conditions: the ratios of shoulder-width-to-body-length and hip-width-to-body-length must exceed predefined thresholds specific to the pigsty environment. Additionally, the triangle formed by the head keypoint and the two shoulder keypoints, as well as the triangle formed by the head keypoint and the left and right hip keypoints, must both be acute triangles. Data satisfying these conditions is deemed valid; otherwise, it is discarded. An example of this filtering is shown in Figure 1.

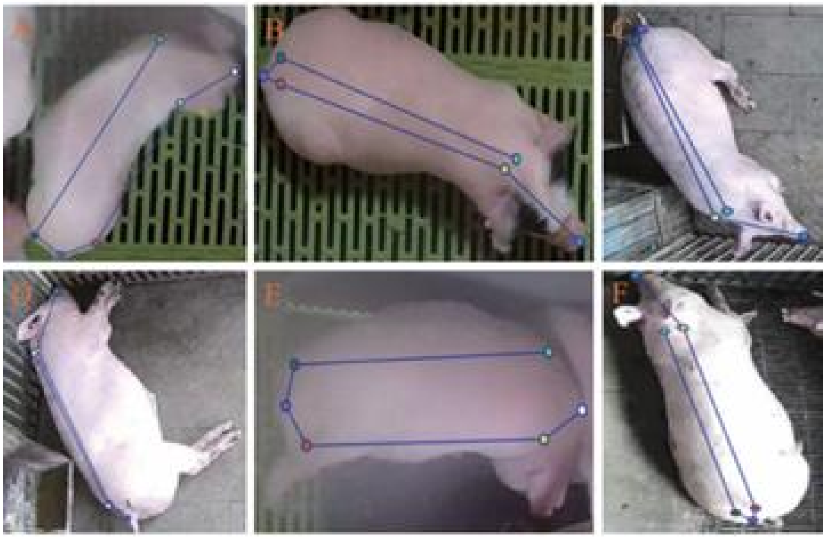


Figure 1 Example of Filtering Abnormal Pig Postures

3.1.2 Abnormal data filtering and calibration

The second stage is abnormal data filtering. Due to factors like motion blur, body dimension data from consecutive frames can exhibit sudden, erroneous spikes. A time-series analysis algorithm is used to identify and filter out these outliers.

To convert pixel measurements into real-world dimensions, the camera's intrinsic (focal length, principal point, distortion coefficients) and extrinsic (position and orientation) parameters are obtained using the MATLAB calibration toolbox. The intrinsic parameters are used to correct for image distortion. By combining both intrinsic and extrinsic parameters, points from the camera's coordinate system are mapped to the world coordinate system, yielding the true body length of the pig (Figure 2).

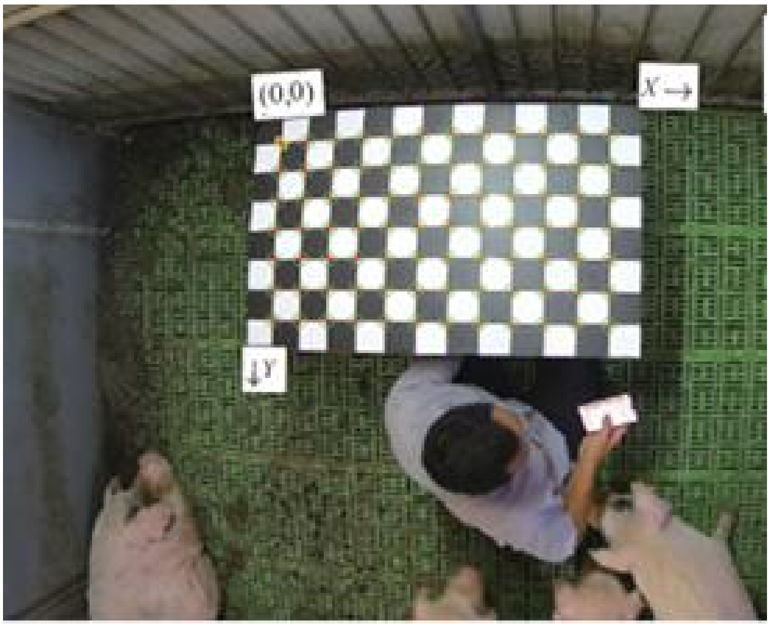


Figure 2 Illustration of Body Dimension Acquisition

3.2 Weight Measurement System

A weighbridge is installed beneath the feeding trough to measure weight, while a camera mounted on a bracket above the trough captures the pig's body dimension data. Pigs tend to stand in a proper posture while eating, making this an ideal time for data collection. To ensure single-pig measurements, an isolation barrier can be used to define the feeding area, and the ROI detection mechanism is employed to discard data from frames containing multiple pigs.

3.3 Temperature Control System

When a deviation exists between the actual ambient temperature and the model-calculated optimal value, the Kalman filter algorithm is used for dynamic optimization and adjustment. The algorithm operates on a "predict-update" cycle, fusing the system model with observation data. The core steps are as follows:

Predict the state:

$$\hat{x}_{k|k-1}=A\hat{x}_{k-1}+Bu_{k-1} \tag{1}$$

Predict the error covariance:

$$P_{k|k-1}=AP_{k-1}A^{T}+Q \tag{2}$$

Calculate the Kalman gain:

$$K_{k}=P_{k|k-1}H^{T}\left(HP_{k|k-1}H^{T}+R\right)^{-1} \tag{3}$$

Update the state estimate:

$$\hat{x}_{k}=\hat{x}_{k|k-1}+K_{k}\left(z_{k}-H\hat{x}_{k|k-1}\right) \tag{4}$$

Update the error covariance:

$$P_{k}=(I-K_{k}H)P_{k|k-1} \tag{5}$$

In this process, the current temperature state is estimated by combining the state transition matrix A, the control input matrix B, and the error covariance matrix. After calculating the Kalman gain Kk, the state estimate is updated, which in turn drives the temperature regulation devices for precise control. Based on the magnitude of the temperature deviation, the system adjusts the power of the heater or the speed of the ventilation fans.

3.4 IoT System Design

The integration of AI empowers the system with capabilities for data learning, pattern recognition, and autonomous decision-making. This project utilizes machine learning algorithms to perform in-depth analysis of real-time data. Combined with the IoT platform, this enables intelligent regulation of the farm's ambient temperature. The data interaction flow is illustrated in Figure 3.

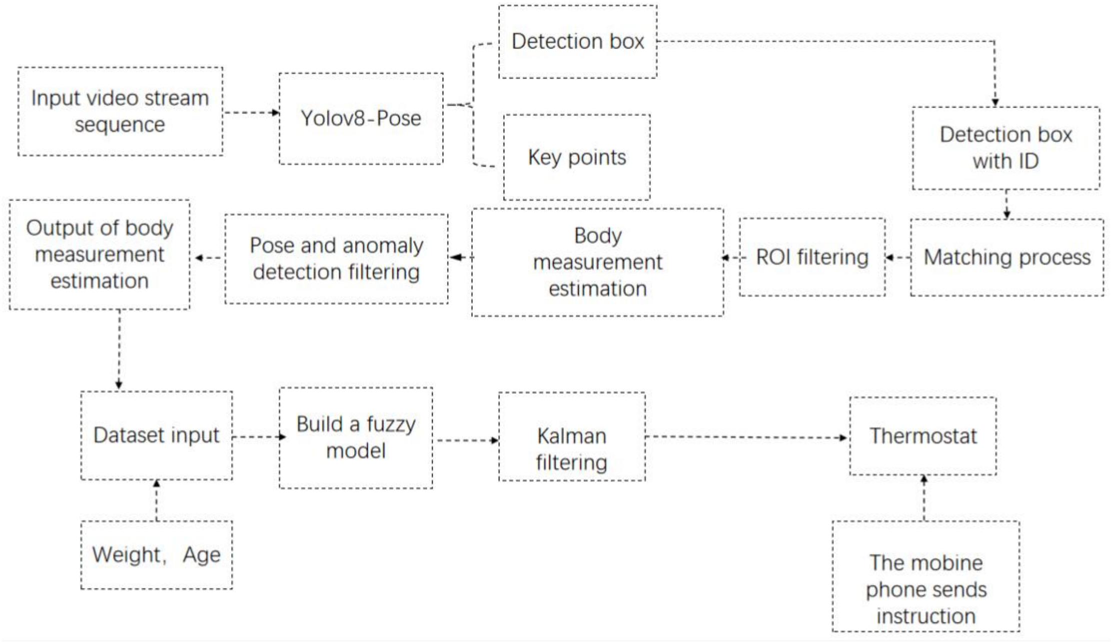

Input video stream sequence
Yolov8-Pose
Detection box
Key points
Detection box with ID
Matching process
ROI filtering
Body measurement estimation
Pose and anomaly detection filtering
Output of body measurement estimation
Dataset input
Weight，Age
Build a fuzzy model
Kalman filtering
Thermostat
The mobine phone sends instruction

Figure 3 AIoT Data Interaction Diagram

4 HARDWARE IMPLEMENTATION

The hardware system is designed for robustness, low cost, and energy efficiency, making it suitable for rural deployment. The key components were selected as follows:

(1) Imaging System: A high-resolution (2448x2048 pixels) GigE industrial camera was chosen for its ability to capture fine details and its reliable data transmission over long distances. It is paired with a white LED bar light source to ensure uniform, stable illumination, minimizing shadows that could interfere with image analysis.

(2) Power Supply: The entire system is powered by monocrystalline silicon solar panels, leveraging a sustainable energy source abundant in rural areas to reduce operational costs and enhance system autonomy.

(3) Control and Communication Core: An STM32C8T6 microcontroller serves as the main controller due to its wide operating temperature range and sufficient peripheral interfaces. For IoT connectivity, the ESP8266 Wi-Fi module is used to transmit data to the cloud platform, enabling remote monitoring and management.

(4) Sensing and Actuation: A CHT11 digital sensor provides accurate real-time temperature and humidity readings. For temperature regulation, a safe, energy-efficient heating element with minimal light emission was selected to avoid startling the pigs or interfering with the camera.

The integrated hardware setup ensures a seamless flow of data from visual capture to environmental actuation, forming a self-contained, intelligent control unit.

5 RESULTS

The system's performance was validated through a series of experiments in a simulated farm environment. The contactless body dimension measurement module demonstrated high accuracy, achieving a Mean Absolute Percentage Error (MAPE) of just 1.5% when compared to manual ground-truth measurements. The posture and anomaly filtering algorithms were critical, successfully filtering out approximately 35% of invalid frames due to unsuitable animal posture or motion blur, thereby ensuring the reliability of the data fed into the control model. In the temperature control evaluation, the system's performance was benchmarked against a traditional On/Off thermostatic controller. When tasked with raising the ambient temperature from 15°C to a target of 24°C, the proposed control mechanism, utilizing a Kalman filter, achieved a smooth and rapid response. It reached the setpoint with a minimal overshoot of less than 0.3°C and maintained the temperature within a highly stable range of ±0.5°C. In contrast, the traditional controller produced significant oscillations, with fluctuations of up to ±2.0°C around the target. This comparison highlights our system's superior stability and energy efficiency. Throughout a 48-hour continuous test, the entire AIoT system, powered by solar panels, operated autonomously and reliably. Real-time data, including pig dimensions and environmental status, was successfully transmitted to the cloud platform with a latency of under 3 seconds, enabling effective remote monitoring and management.

6 CONCLUSION

This project successfully demonstrates an intelligent temperature control system for pig farming by integrating a contactless body dimension detection system with real-time environmental feedback. The innovative use of AI-driven computer vision, fuzzy logic, and IoT technologies provides a novel solution for automated environmental management in livestock farming. The emphasis on system integration, safety, and environmental adaptability, particularly through the use of solar power, enhances the practical utility and sustainability of the solution. This work offers a valuable and reliable technological advancement for farmers, contributing to the ongoing progress and development of the modern livestock industry.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was financially supported by “Innovation and Entrepreneurship Training Program for Jiamusi University Students”(S202510222140) fund. Team members include Shixin Qi, Xiaojun Bai, Jiayong Wu, Han Zhang.

REFERENCES

[1] China Animal Agriculture Association. China's Swine Industry Development Report. Beijing: China Agriculture Press, 2023.
[2] Wang Xiaopin, He Wei, Guo Yangyang. Research progress on the application of machine vision in pig farming. Journal of Agricultural Engineering, 2024, 40(3): 1-12.
[3] Ultralytics. Yolov8: Real-Time Object Detection and Instance Segmentation. 2023. https://github.com/ultralytics/ultralytics.
[4] Hidra N, Lehmad M, Doubabi S, et al. Design and implementation of an intelligent remote monitoring and control system for enhancing hybrid solar-electric dryer (HSED) performance. Drying Technology, 2025, 43(11-12): 1855-1878.
[5] Hao Y, Li J, Wang C. A 3D point cloud analysis system for livestock body measurement. Computers and Electronics in Agriculture, 2017, 140: 213-221.
[6] Bouarroudj K, Babaa F, Touil A. IoT-based monitoring and control for optimized plant growth in smart greenhouses using soil and hydroponic systems. Internet of Things, 2025.
[7] Liu Tonghai, Zhang Yong, Li Qingda. 3D model construction and body measurement of pigs based on irregular triangular network. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(8): 256-261.
[8] Velpula R M, Veluvolu R V. A Novel Intelligent Controller for Enhancing the Dynamic Performance of an SVM-DTC Based Induction Motor Drive. Arabian Journal for Science and Engineering, 2025.
[9] Pimpalkar R. A smart solar PV monitoring system using internet of things (IoT). Concurrent Engineering, 2025, 33(1-4): 50-60.
[10] S I S, Senthilkumar T, Manikandan G, et al. Development of a smart remote-controlled system for four-wheel paddy transplanter with anti-collision safety system and vision-assistance system for precision agriculture. Results in Engineering, 2025.
[11] Prabakar D, Meenalochini P, A R B, et al. A hybrid approach based internet of things assisted power monitoring system for smart grid. Analog Integrated Circuits and Signal Processing, 2025, 125(2): 30-30.
[12] Boretti A. Advanced Battery Thermal Management: A Review of Materials, Cooling Systems, and Intelligent Control for Safety and Performance. Energy Storage, 2025, 7(7).

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3103

SEMANTIC PRIVACY RISKS IN SOCIAL TRAJECTORY PUBLICATION

ZhenZhen Wu, YanFei Yuan*
College of Cyber Security, Tarim University, Alar 843301, Xinjiang, China.
Corresponding Author: YanFei Yuan, Email: yanfeiyuan@taru.edu.cn

Abstract: The publication of trajectory data in mobile applications raises significant privacy concerns for users. When combined with behavioral pattern analysis, the semantic similarity of published trajectories can be exploited by attackers to infer users' travel motivations, posing substantial risks to personal privacy. In this paper, we simulate such attacks by proposing an observation-based algorithm to infer user travel behavior and develop a corresponding privacy risk quantification mechanism. Extensive experiments on real-world datasets validate the effectiveness of the proposed risk quantification approach, providing a foundation for the further development of semantic privacy protection schemes.

Keywords: Mobile application; Behavioral semantics; Privacy inference; Risk quantification

1 INTRODUCTION

The rapid advancement of cloud computing, IoT, and wireless communication technologies has accelerated the growth of mobile communication services [1]. These services are now widely adopted across diverse domains including social entertainment, intelligent transportation, military and defense, as well as government and business operations [2]. To deliver functional and personalized services—such as point-of-interest searches, navigation, and weather forecasts—mobile applications often require users to share their location trajectories [3]. However, the publication of such trajectory data raises significant privacy concerns among users [4].

In response, a variety of location privacy protection schemes have been developed. These include traditional techniques such as anonymization, generalization, obfuscation, and perturbation [5]. For instance, anonymization blends real locations with fake ones [6]; generalization and obfuscation replace precise points with broader regions [7]; and perturbation introduces noise into actual location coordinates [8]. Additionally, geographic indistinguishability applies differential privacy to constrain the distinguishability between real and synthetic locations [9]. Despite their utility, these methods primarily safeguard user privacy within the geographic domain.

Nevertheless, privacy in mobile travel encompasses not only geospatial location protection but also behavioral privacy in the semantic dimension—such as inferring a user's travel motivation. To address this, Dai et al. introduced a semantic generalization method that reduces the semantic similarity between reconstructed trajectories and the user's actual behavior [10]. Our prior work [11] enhanced this approach by adaptively adjusting generalization strength based on the user's access roles at different locations. More recently, we further refined semantic generalization by incorporating behavioral patterns to govern privacy sensitivity-aware strength [12].

These efforts highlight the need for a clear criterion to quantify the appropriate degree of semantic generalization in location privacy protection. Without such a guideline, published trajectories may either reveal user travel motivations due to excessive semantic similarity, or degrade service quality due to overly distorted semantic information.

To bridge this gap, this paper proposes a quantification mechanism that assesses the risk of privacy leakage arising from the semantic similarity of published trajectories. This mechanism offers a quantitative criterion to guide semantic generalization in location privacy protection. The main contributions of this work are as follows.

2 SEMANTIC SIMILARITY

We first represent and quantify the semantic similarity between location functional attributes to support subsequent behavioral inference based on published trajectories.

2.1 Trajectory Data Preprocessing

Location data generated from user mobility, often captured by global navigation satellite systems (GPS, GLONASS, Galileo, BDS), form geographic trajectories (Geo-Trajectories). These trajectories can be formally represented as a sequence of visited points, Geo_Traj={vi}Geo_Traj={vi}, where each point vi=(ti,li), vi=(ti,li) consists of a timestamp ti and a geographic coordinate li=(lati, longi, li=(lati, longi).

The check-in locations within a user's mobile trajectory can be broadly classified into two types: moving points and staying points. Moving points describe the path, speed, and distance of travel. In contrast, the purpose of a user's social travel is often revealed by the functional attributes of the places where they stay—for instance, sleeping at home at night, working at an office during the day, or shopping at a mall. This paper focuses exclusively on stay locations and the transitions between them, disregarding moving points and the specific geographic paths taken by the user [3]. We annotate the behavioral semantics of a user by analyzing the functional attributes of their identified stay locations.

User social travel typically involves two forms of staying: remaining at a specific location for a prolonged period, or wandering within a defined area. Based on these patterns, we formally define the concept of a stay point and describe a method for extracting these points from user travel trajectories.

2.2 Behavioral Semantic Annotation

We infer the semantic functional attributes of stay points by leveraging open-source web data related to the corresponding locations or regions. This process facilitates the discovery of a user's travel purpose at each stay point and reveals the underlying behavioral semantics of their social trips. The assignment of such semantic properties to user -visited locations is commonly referred to as semantic annotation.

One approach to semantic annotation involves analyzing the functional properties of Points of Interest (POIs) near a user's stay location using map applications—such as Google Earth, Baidu Map, or OpenStreetMap [11]. In this method, typical semantic categories are selected and applied to label the stay location. Common semantic selection algorithms include proximity-based principles, quantity-first selection, and TF-IDF.

Alternatively, open-source user-generated content from social network platforms—such as Foursquare, Instagram, Sina Weibo, Meituan, and RenRen—can be utilized [7,13]. This includes travel check-ins, location-related tweets, comments, reviews, and ratings. Through semantic analysis of such social data, the functional attributes of a user's stay location can be effectively derived.

These procedures enable the transformation of raw geographic trajectories into behavior-oriented sequences composed of semantic attributes from stay points. This structured representation supports the characterization of users' social mobility patterns and enables in-depth behavioral semantic analysis.

2.3 Hierarchical Semantic Similarity Architecture

The semantic-functional properties of Points of Interest (POIs) in the real world typically exhibit a hierarchical similarity structure. For instance, junior high schools, high schools, and universities all fall under the broader category of educational institutions, while also encompassing more specific subtypes. Similarly, Chinese and Western cuisine belong to the food category; shopping malls and supermarkets can be grouped under shopping; and activities such as dining, shopping, KTV, bars, and cinemas are often classified as entertainment.

Building upon these observations, we construct a hierarchical tree structure to characterize semantic similarities among POIs. First, the specific semantic functional attributes of POIs are treated as leaf nodes in the tree. We then perform semantic generalization and clustering at the finest possible granularity, grouping entities such as KTV, bars, cinemas, and supermarkets into progressively broader categories. Through iterative application of this process, a hierarchically organized semantic tree is ultimately established.

This semantic tree captures similarity relationships among semantic attributes through its layered organization, analogous to kinship structures in human society. It provides a powerful technical framework for conducting semantic analysis in mobile computing scenarios.

2.4 Association Representation of Semantic Similarity

To characterize the mapping relationships of behavioral semantics between synthetic and actual trajectories, we quantify location semantic similarities using the constructed semantic tree. This quantification establishes the basis for evaluating privacy leakage risks in trajectory publication.

The semantic similarity quantification process proceeds as follows: First, we select an arbitrary leaf attribute as the starting point and traverse upward through the semantic tree layer by layer. During this backtracking process, we cluster the newly encountered attributes (Aτ) from the leaves of each intermediate node. The mapping probability from the departure attribute to each newly traversed attribute, denoted as Ri(oj) is calculated to be inversely proportional to the size of Aτ and decays with the number of iterations.

This upward traversal continues until the root node is reached, establishing semantic similarity association probabilities from the departure node to all other attributes. Finally, we systematically replace the departure attribute and repeat the entire traversal process until every attribute has served as a starting point. Through this comprehensive procedure, we obtain complete semantic similarity associations between all attribute pairs.

3 BEHAVIORAL INFERENCE ASSOCIATED WITH SEMANTIC SIMILARITY OF PUBLISHED TRAJECTORY

We represent the stochastic process of user social mobility following the principle of HMM, and reveal the risks of privacy leakage induced by the semantic similarity of published trajectories to users' real behaviors.

3.1 User Behavioral-Pattern Construction

Following data preprocessing and behavioral semantic annotation, we obtain the user's behavioral trace in mobile scenarios. To model these traces, we construct behavioral patterns of user social travel using a supervised learning approach based on the Markov Chain principle.

Two key challenges must be addressed: first, user travel behavior is closely tied to their current life-state; second, it exhibits strong temporal dependencies. For instance, users typically return home at night, work during daytime hours, and engage in leisure activities during holidays. Traditional Markov models fail to capture these complex spatio-temporal-life-state associations, focusing solely on spatial transitions [14].

To overcome these limitations, we enhance the Markov model by constructing a time partition-based extension matrix M~ {aij} and multiple life-state-specific matrices AM~ Mls. This extension incorporates temporal dimensions into the spatial transitions, representing each transition as aij~ ((ti,si)→(tj,sj)) rather than simply (li,lj). This formulation captures both the user behavior and its temporal context. Additionally, by maintaining separate matrices for different life-states, we prevent interference between distinct behavioral patterns during characterization.

3.2 Behavioral-Semantic Mapping Associations between Published and Actual Trajectories

Semantic similarity constitutes the fundamental basis for mapping associations Ri(oj) between user behaviors in published and actual trajectories. As established in Section 3.1, we characterize these mapping associations by incorporating the computed semantic similarities.

The association probabilities Ri(oj) between any two attributes are derived through backtracking operations on the semantic tree. We construct a two-dimensional observation matrix BM={Ri(oj)} for our Hidden Markov Model (HMM) using these probabilities. This matrix effectively represents the semantic associations between published observation attributes oj and potential actual behavioral semantics si.

3.3 Social-Mobility Stochastic Processes

Building upon the constructed HMM, we model the stochastic process of invisible user social mobility in mobile scenarios. We define an HMM-based forward variable αα and specify its computational method, describing its evolution process. During evolution, target forward variables are computed and intermediate results are stored in a matrix structure. These stored results subsequently support efficient behavioral inference while reducing computational overhead.

3.4 Observation-based Behavioral Inference

In mobile social applications, attackers may observe users' shared location trajectories. Through behavioral pattern analysis of published trajectories, they can infer users' upcoming behaviors. This section simulates such inferential attacks to quantify the mobile privacy leakage risk associated with trajectory publication.

As published trajectories consist of sequentially shared access locations, we dynamically quantify privacy leakage risk by characterizing the attacker's inference probability before and after each location observation. This iterative approach ultimately achieves comprehensive privacy risk assessment for the complete published trajectory.

Let $P^-(s)$ represent the attacker's prior inference probability about the user's actual behavior before observing a released location, and $P^+(s \mid z)$ denote the posterior probability after observation. We further transform these formulae using the characterized HMM-based social mobility and the defined forward variable:

$\Delta = P^+(s|z) - P^-(s)$

$P^+(s|z) = p(st|o1,\cdots,ot-1,ot= z,\lambda)$

$=p(o1,\cdots,ot-1,ot= z,st|\lambda)/p(o1,\cdots,ot-1,ot= z|\lambda)$

$P^-(s) = p(st|o1,\cdots,ot-1,\lambda)$

$=p(o1,\cdots,ot-1,st|\lambda)/p(o1,\cdots,ot-1|\lambda)$

4 EXPERIMENTAL EVALUATION

4.1 Datasets and Setup

Mobile Dataset: We utilized the real-world Geolife dataset from the Microsoft Research Asia project led by Yu Zheng's group to evaluate the performance of BSPri. This dataset captures daily life trajectories of 182 users over a five-year period, primarily within Beijing. Our experiments focus on mobility data inside Beijing's Sixth Ring Road.

Open-source Libraries: Through Baidu Map API, we collected public POIs within the target area to construct a POI Library containing 90,494 entries. Their semantic attributes were extracted to form an open-source Semantic Library with 44 standardized categories. Using time/distance thresholds of 1min/10m and 30min/100m, we identified 9,495 stay points from the Geolife dataset, establishing a Staying-points Library as personal mobility data.

Privacy-preservation Setup: Our semantic generalization mechanism for trajectory reconstruction operates through backtracking on the semantic tree during similarity characterization. This process uses leaf attributes of intermediate nodes reached after backtracking as candidate anonymous semantic types, with backtracking levels indicating the degree of semantic generalization. We conducted experiments to quantify behavioral semantic leakage risks across different generalization levels.

4.2 Experimental Results of Privacy Risks

We deployed inference algorithms in scenarios both with and without life-state differentiation to assess travel behavior leakage risks during weekdays and holidays. The following figures presents privacy leakage metrics through both dynamic progression and average values.

Comparative analysis of Figures 1 and 2 reveals two key findings: First, with one-level backtracking, consistently high metric values indicate that weak semantic generalization significantly increases attackers' inference probability, substantially exposing travel privacy. Second, while three-level backtracking effectively constrains overall metric values, sporadic peaks suggest the necessity of excluding high-similarity attribute types in privacy protection design - a consideration we incorporate in subsequent work.

Figure 3 demonstrates that average inference probabilities decrease substantially with increasing backtracking levels, confirming semantic generalization as an effective countermeasure against behavioral-semantic inference attacks. Although non-life-state-differentiated curves closely resemble weekday patterns, holiday curves exhibit lower inference probabilities and more pronounced variation trends under semantic generalization, reflecting users' diverse behavioral patterns during holidays.

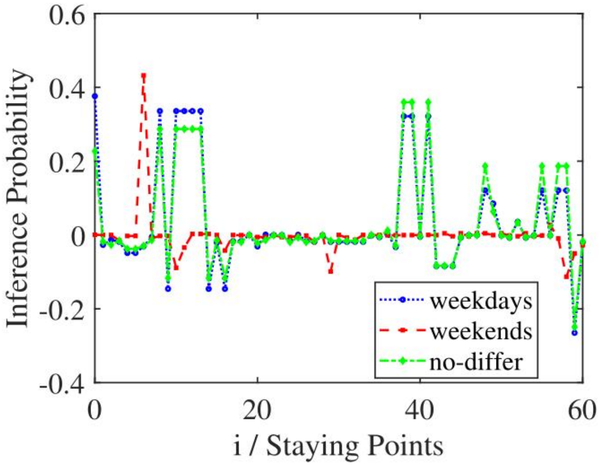


Figure 1 Dynamics of Inference Probability in Different Scenarios with bk = 1

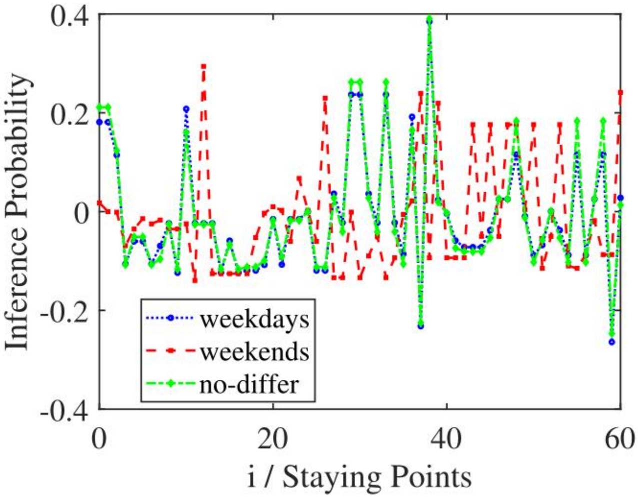


Figure 2 Dynamics of Inference Probability in Different Scenarios with bk = 3

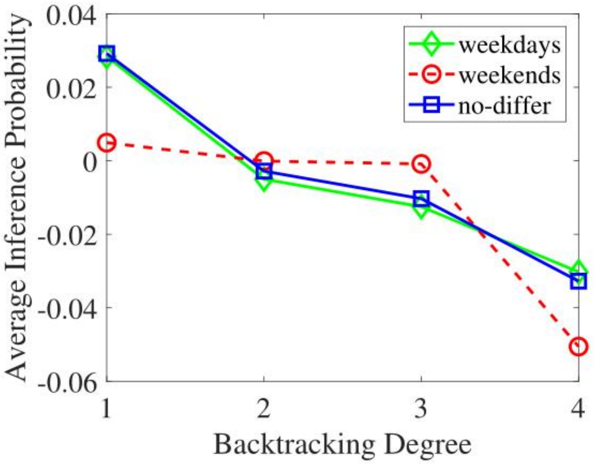


Figure 3 Average Inference Probability in Different Scenarios

5 CONCLUTION

Addressing the privacy concerns in mobile trajectory publication, this paper presents a privacy risk quantification mechanism. We simulate inference attacks by developing an observation-based behavioral inference algorithm, which characterizes the privacy leakage risk stemming from the semantic similarity between published trajectories and users' actual behaviors. Extensive experiments on real-world datasets validate our approach by demonstrating the leakage risks under varying degrees of semantic generalization, thereby providing a foundation for designing subsequent semantic privacy protection schemes.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Qiu G, Tang G, Li C, et al. Differentiated location privacy protection in mobile communication services: A survey from the semantic perception perspective. ACM Computing Surveys(CSUR), 2023, 56(3): 1-36.

[2] Jiang H, Li J, Zhao P, et al. Location privacy-preserving mechanisms in location-based services: A comprehensive survey. ACM Computing Surveys(CSUR), 2021, 54(1): 1-36.

[3] Zheng Y. Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology(TIST), 2015, 6(3): 1-41.

[4] Jin X, Zhang R, Chen Y, et al. Dpsense: Differentially private crowdsourced spectrum sensing. In Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security(CCS), Vienna, Austria, 2016: 296-307.

[5] Primault V, Boutet A, Mokhtar SB, et al. The long road to computational location privacy: A survey. IEEE Communications Surveys and Tutorials, 2019, 21(3): 2772-2793.

[6] Zhang J, Li C, Wang B. A performance tunable cpir-based privacy protection method for location based service. Information Sciences, 2022, 589: 440-458.

[7] Zhao W, Zhou N, Zhang W, et al. A probabilistic lifestyle-based trajectory model for social strength inference from human trajectory data. IEEE Transactions on Information System, 2016, 35(1): 1-28.

[8] Song C, Raghunathan A. Information leakage in embedding models. In Proceedings of the 27th ACM SIGSAC Conference on Computer and Communications Security(CCS), Gather.town, virtual platform, 2020: 377-390.

[9] Andrés ME, Bordenabe NE, Chatzikokolakis K, et al. Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, 2013: 901-914.

[10] Dai Y, Shao J, Zhang D. Personalized semantic trajectory privacy preservation through trajectory reconstruction. World Wide Web, 2018, 21(4): 875-914.

[11] Qiu G, Guo D, Shen Y, et al. Mobile semantic-aware trajectory for personalized location privacy preservation. IEEE Internet of Things Journal, 2020, 8(21): 16165-16180.

[12] Qiu G, Tang G, Li C, et al. Behavioral-semantic privacy protection for continual social mobility in mobile internet services. IEEE Internet of Things Journal, 2024, 11(1): 462-477.

[13] Yang C, Sun M, Zhao WX, et al. A neural network approach to jointly modeling social networks and mobile trajectories. IEEE Transactions on Information System, 2017, 35(4): 1-28.

[14] Phan N, Wang Y, Wu X, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, Arizona USA, 2016: 1309-1316.

Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3104

ADVANCES IN INTELLIGENT ROCK IMAGE RECOGNITION BASED ON CONVOLUTIONAL NEURAL NETWORKS

He Ma
Yellow River Engineering Consulting Co., Ltd., Zhengzhou 450003, Henan, China.
Corresponding Email: mahe026@163.com

Abstract: Lithology identification is a fundamental task in resource exploration and engineering geology, yet traditional methods face bottlenecks such as low efficiency and high subjectivity. In recent years, intelligent rock image recognition techniques based on convolutional neural network (CNN) have demonstrated remarkable advantages. This paper systematically reviews the research progress of CNN in intelligent rock image recognition and explores their application potential and technical challenges in this field. First, the basic architecture and working principles of CNN are introduced, including the synergistic interactions among convolutional layers, pooling layers, and fully connected layers. Subsequently, the criteria for model selection and optimization pathways in rock image recognition are analyzed, covering task-specific model adaptation strategies and multi-model comparative evaluation and selection strategies. Additionally, the roles of data augmentation strategies, resolution enhancement techniques, and model architecture innovations in improving model performance are discussed. Finally, this paper summarizes the limitations of current research and proposes future research directions, aiming to provide theoretical support and practical guidance for overcoming existing technical bottlenecks.

Keywords: Convolutional neural network; Rock image; Lithology identification; Identification mode; Optimization path

1 INTRODUCTION

Lithology identification serves as a core foundational task in resource exploration and engineering geology. Precise identification results not only provide critical data support for deep mineral exploration and hydrocarbon reservoir evaluation but also offer a scientific basis for design optimization and construction safety in major engineering projects, such as mining, tunneling, and hydraulic infrastructure development [1]. Although traditional methods—including macroscopic observation, thin-section identification, and laboratory analysis—have been widely applied [2-7], these workflows are characterized by high labor intensity, substantial costs, prolonged duration, and significant subjectivity [8 -11].

In recent years, the rapid advancement of artificial intelligence has provided technical support for the intelligent detection, classification, and segmentation of rock imagery, offering a novel pathway to mitigate the excessive reliance on expert experience inherent in traditional lithology identification. Current intelligent recognition technologies based on rock images are primarily categorized into two distinct paradigms according to their automation levels: traditional machine learning (ML)-based classification and deep learning (DL)-based intelligent recognition [12]. Regarding ML approaches, Marmo and Amodio utilized a multilayer perceptron neural network to identify mudstones and clastic rocks based on 23 distinct features [13], such as grayscale percentages and edge pixel counts. Similarly, Chatterjee applied support vector machines (SVM) to limestone identification using 40 attributes involving color, texture [14], and morphology; Cheng and Yin employed SVM with 13 spatial and textural features to classify four rock types [15]; and Izadi et al. conducted intelligent identification of 23 igneous rocks using artificial neural networks (ANN) based on 12 color parameters [16]. Although ML-based classification can effectively differentiate lithologies, it necessitates the manual extraction of features for rock delineation. Consequently, these techniques often suffer from diminished efficiency and accuracy when processing large-scale or diverse datasets [17-18].

Deep learning algorithms, such as CNNs, utilize unique hierarchical abstraction mechanisms to adaptively extract multi -scale spatial features directly from raw pixel data. This establishes an end-to-end learning framework that eliminates the need for manual feature selection, thereby further attenuating subjectivity in lithology identification [12, 19-21]. For instance, Zhang and Li constructed an intelligent identification model using Inception-V3 for rocks such as granite [20], phyllite [22], and breccia, achieving a classification accuracy exceeding 90% on the test set. Feng and Gong developed a Siamese CNN model based on AlexNet for 28 rock types, reporting a test accuracy of 89.4%. Furthermore, Ran and Xue successfully classified granite, limestone [23], conglomerate, sandstone, shale, and mylonite using a custom RTCNN architecture, attaining an accuracy of 97.96%. Building upon these foundations, researchers have deployed such models onto mobile devices, enabling rapid in-situ lithology identification for field geological surveys [24-26].

Despite the immense potential of CNNs in the domain of rock image recognition, practical application remains constrained by multi-dimensional technical bottlenecks. First, the proliferation of CNN architectures has intensified the dilemma of model selection [27-31], where traditional trial-and-error methods struggle to adapt to the multi-scale feature requirements of rock imagery. Second, data scarcity and the "annotation paradox" restrict model generalization; rock sampling is limited by geological conditions, and high-precision labeling relies on professional geologists, resulting in a paucity of large-scale annotated datasets. Furthermore, class imbalance and inter-class similarity exacerbate decision risks. The natural distribution of rock types is uneven (e.g., sedimentary rocks significantly outnumber metamorphic rocks), and subtle textural differences between similar lithologies (such as argillaceous sandstone versus sandy mudstone) lead to persistent misclassification rates [8, 10, 32-34]. Finally, image degradation caused by environmental interference—such as uneven lighting and surface contamination during field acquisition—diminishes feature validity, as low-resolution images fail to preserve critical textural details [35-40]. Collectively, these challenges severely impede the generalization capability and engineering applicability of AI models.

Against this background, constructing a robust intelligent rock identification framework that transcends the synergistic constraints of data, models, and environment has become a focal point for both academia and industry. This paper systematically reviews the research progress of CNN-based rock image identification, focusing on model selection criteria and optimization pathways. The objective is to provide theoretical underpinning for overcoming existing technical bottlenecks.

2 INTRODUCTION TO CONVOLUTIONAL NEURAL NETWORKS

Since its inception in the 1980s, the CNN has established itself as a cornerstone tool in the field of image recognition. Its canonical architecture comprises convolutional layers, pooling layers, and fully connected layers, which synergize to realize end-to-end learning from raw input images to final classification results (Figure 1). The convolutional layer, acting as the backbone of the CNN, extracts local features through convolution operations between kernels and local receptive fields of the image. This mechanism captures critical low-level information, such as edges and textures, providing a fundamental basis for subsequent tasks. Interspersed between successive convolutional layers, the pooling layer performs down-sampling to progressively diminish the spatial dimensionality of feature representations. This process minimizes the number of model parameters and computational load while serving to mitigate overfitting. Positioned at the terminus of the architecture, the fully connected layer integrates and maps the extracted features to derive high-level semantic information. By establishing dense connectivity where each neuron links to all neurons in the preceding layer, it applies non-linear transformations via activation functions to output the final classification results.

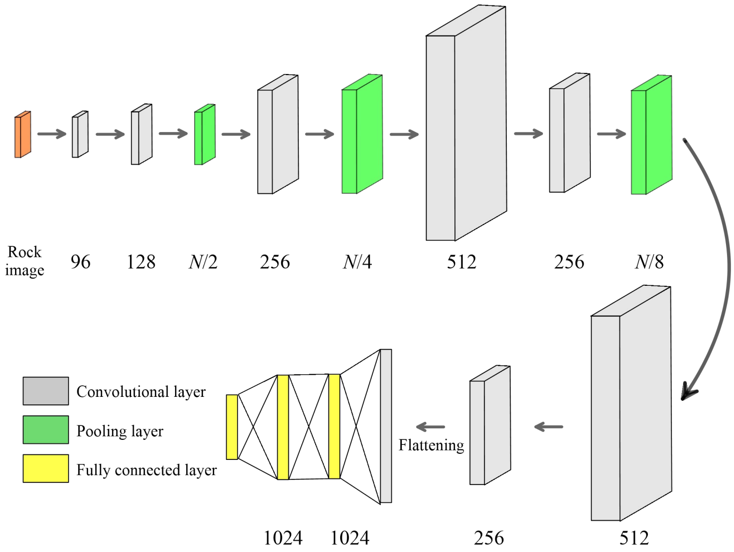


Figure 1 Architectural Diagram of the Convolutional Neural Network

Driven by the continuous evolution of the ILSVRC competition in recent years, numerous deep learning architectures demonstrating superior image classification performance have emerged, including AlexNet [27], VGG [31], Inception-V3 [29], ResNet [30], and Xception [28]. The structural and algorithmic innovations within these models have solidified the foundation for deep learning applications in image classification. Specifically, relative to conventional neural networks, AlexNet incorporated Dropout and ReLU activation functions, significantly accelerating training convergence and enhancing performance (Figure 2a). VGG16 deepened the network architecture by employing small-sized 3×3 convolution kernels while maintaining manageable parameter counts (Figure 2b). Inception-V3 introduced auxiliary classifiers and label smoothing techniques to effectively alleviate the vanishing gradient problem and ensure gradient stability, thereby improving model generalizability and preventing overfitting (Figure 2c). ResNet, through the introduction of residual units, optimized parameter configuration while achieving comparable accuracy; this design enables the model to attain desired performance levels with fewer iterations, substantially boosting training efficiency (Figure 2d). Finally, Xception adopted depthwise separable convolutions to drastically reduce the number of model parameters while preserving high performance (Figure 2e).

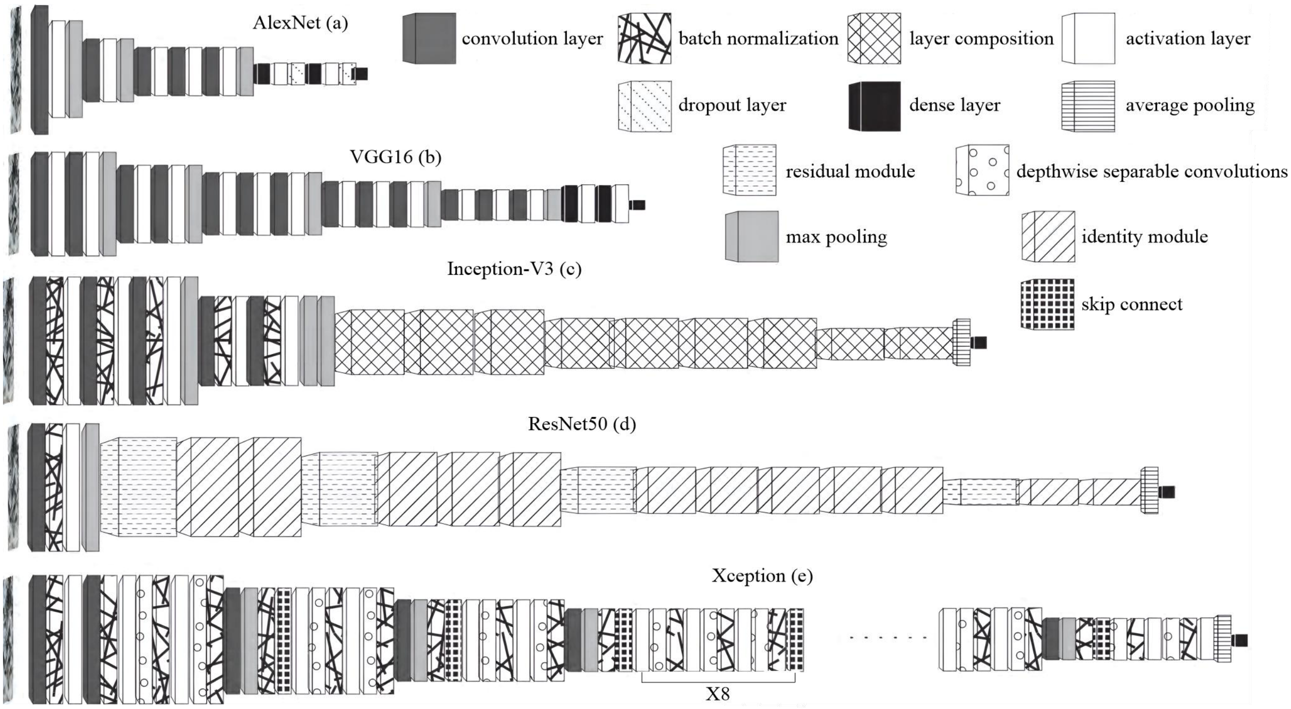


Figure 2 Schematic Diagram of a Typical Convolutional Neural Network Architecture

3 OPTIMIZATION PARADIGMS FOR CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES

In the realm of deep learning, the relationship between the complexity of model architectures and their feature extraction capabilities exhibits non-linear characteristics, a phenomenon particularly pronounced in research on intelligent rock image recognition. Classical theory posits that increasing network depth facilitates the extraction of high -order abstract features, thereby enhancing image recognition performance [30]. However, recent investigations indicate that when the number of layers exceeds a critical threshold, models may experience accuracy saturation or even performance degradation [21, 36]. This finding has been validated within the domain of rock lithology identification: comparative studies involving AlexNet, ResNet, Inception, and VGG [22], as well as evaluations across AlexNet, VGG16, ResNet50, and Xception [41], and layer-depth contrast experiments with ResNet-50/101/152 [38], have all demonstrated that models with lower structural complexity often exhibit superior classification performance. These results suggest that relying solely on network depth as a selection criterion has limitations, necessitating a comprehensive trade-off between computational efficiency and engineering applicability [42]. Currently, academia primarily adopts two optimization strategies: (1) adaptation based on model characteristics, and (2) multi-model comparative selection. The advantages of these approaches in practical applications are discussed below.

3.1 Adaptation Strategies Based on Model Characteristics

In research on CNN-based intelligent rock image recognition, the selection of model architectures often involves multi-dimensional technical considerations. For instance, Bai and Yao [8] pioneered the use of the Inception-V3 architecture for constructing a rock image transfer learning model, a decision primarily grounded in the model's historical success in general image recognition tasks. As research deepened, Hu and Ye [40] proposed a dual criterion for model selection: prioritizing network architectures like ResNet-50, which possess moderate parameter counts and high training efficiency, while ensuring recognition accuracy; this approach effectively balances computational resource consumption with model performance. Notably, the preference for ResNet models by Yang and Xiong and Wang and Liu was largely driven by the architecture's ability to mitigate issues such as vanishing gradients [36, 43], exploding gradients, and network degradation during training. In contrast, distinct technical priorities exist among different research groups. For example, despite acknowledging the risks of saturation and degradation associated with increased depth, some researchers selected the ResNet-101 model [21]. Systematic experiments revealed that for complex lithology recognition tasks, extending network depth to ResNet-101 yielded significantly better feature representation capabilities than ResNet-50. These studies indicate that in the field of intelligent rock image recognition, model selection requires not only an assessment of baseline performance metrics but also a comprehensive integration of specific geological scenario requirements, computational efficiency, network depth, and gradient optimization.

3.2 Multi-Model Comparative Selection Strategy

Comparative analysis represents another pivotal methodology for selecting appropriate architectures. As illustrated in Table 1, Alzubaidi and Mostaghimi [34] systematically evaluated three typical CNN models—ResNeXt-50, Inception-v3, and ResNet-18. They discovered that ResNeXt-50 demonstrated significant superiority in core image recognition tasks, improving accuracy by 10 percentage points compared to Inception-v3 while reducing training time by 43%; consequently, it was selected as the core architecture for their intelligent lithology identification system. This finding resonates with the large-scale model screening conducted by Ma and Ma [38], who initially filtered 13 mainstream models (including VGG16, MobileNet, and Xception) before conducting an in-depth comparison of five candidates, such as ResNet50 and Inception-v3. Their experimental results indicated that ResNet50 achieved the optimal balance between model complexity and recognition performance, establishing it as the underlying pre-trained model for lithology identification. It is worth noting that the superiority of the ResNet series has been validated across multiple independent studies. Ren and Zhang [44] compared current mainstream deep learning algorithms—including AlexNet, VGGNet, GoogleNet, and ResNet—in rock and mineral sample identification tasks, similarly finding that the ResNet50 model yielded the highest recognition accuracy, leading to its selection as the foundational model.

However, recent scholarship has revealed the scenario-dependent nature of model selection. Comparative experiments by Li targeting fresh rock section images demonstrated that the myDenseNet-all model [9], based on the DenseNet architecture, achieved optimal performance on the test set with an F1-score of 89.84% and an accuracy of 94.48%. These metrics exceeded those of the improved myResNet-all model by 2.01 and 1.72 percentage points, respectively, suggesting that dense connectivity structures possess a stronger capacity for capturing the features of fresh sections. This discrepancy was foreshadowed in early research by Feng and Gong [22], where comparative experiments showed that the accuracy (0.66) and F1-score (0.77) of ResNet-vl-50 were significantly lower than the accuracy (0.79) and F1-score (0.87) achieved by AlexNet. Moreover, Zhang and Tang [32] further validated this pattern when constructing a centimeter-scale intelligent identification system for continental shale lithology. They found that EfficientNet-B0 surpassed ResNeXt-50 in both Top-1 (60%) and Top-5 (95%) accuracy, as its compound scaling strategy was better adapted to the microstructural features of centimeter-scale shale thin sections.

A comprehensive analysis indicates that current research on intelligent rock image recognition exhibits a dual development trend characterized by "benchmark model selection + scenario-based adaptation." Although multi-model comparisons can identify high-precision architectures for specific tasks, significant disparities remain in the performance of identical or similar models across different rock image datasets.

Table 1 Comparative analysis of Convolutional Neural Network in Research Citations

Tested model	Optimal model	References
ResNeXt-50, Inceptoin-v3, ResNet-18	ResNeXt-50	[34]
VGGl6, VGG19, MobileNet, AlexNet, LeNet, ZF_Net, ResNet18, ResNet34, ResNet50, Inception-V3, ResNet152, ResNet101, Xception	ResNet50	[38]
AlexNet, VGGNet, GoogleNet, ResNet	ResNet50	[44]
DenseNet121, EfficientNet-BO, EfficientNet-B8, MobileNet-v2, MVIT-v2, ResNext50	EfficientNet	[32]
AlexNet, ResNet-vl-50, Inception-V2, VGG-19	AlexNet	[22]
VGG, ResNet, DenseNet	myDenseNet-all	[9]

4 OPTIMIZATION PATHWAYS FOR INTELLIGENT ROCK IMAGE IDENTIFICATION MODEL PERFORMANCE

Building upon diverse selection criteria, scholars have established multiple architectural frameworks for intelligent rock image identification. However, existing research indicates that model accuracy on test sets generally possesses room for optimization. The primary constraints can be categorized into three aspects: first, the insufficient scale of training data fails to meet big data volume requirements, significantly impacting model generalization capabilities [8, 10, 20, 33, 36, 39, 40, 44-49]; second, low image resolution leads to inadequate feature extraction, resulting in identification errors for lithologies with similar mineral compositions or macroscopic characteristics [8, 10, 33, 36, 38-40, 43, 50]; and third, the discriminative capability of model architectures requires enhancement through further optimization or the adoption of more refined models [10, 22, 37, 38, 43, 44, 46, 47, 51]. Based on this analysis, the following sections systematically review the latest research progress across three dimensions: data augmentation strategies, resolution enhancement technologies, and model architecture innovation**.**

4.1 Data Augmentation Strategies: From Limited Samples to Robust Feature Learning

The core of Convolutional Neural Networks lies in learning intrinsic feature representations of rock images via a data-driven approach. Consequently, the scale and quality of training data directly influence the model's ability to generalize lithological features. When sample sizes are insufficient, models are prone to overfitting local features, leading to a significant decline in lithological discrimination capability [20]. Currently, two primary data augmentation paradigms are employed [10, 32, 45, 48]: (1) Physical augmentation, which involves acquiring incremental raw data through multi-angle rock sampling (e.g., 3D rotational scanning, multi-spectral imaging); and (2) Digital augmentation, which expands datasets using traditional methods such as geometric transformations (flipping/rotation/translation), photometric adjustments (brightness/contrast perturbation), and noise injection (Gaussian/Salt-and-Pepper noise).

Regarding intelligent identification models based on Inception-V3, Zhang and Li observed that classification probability values for two granite images and one breccia image in the test set were below 70% [20]. By augmenting training data through local image cropping, model accuracy rose to over 85%. This study suggests that data augmentation targeting local textural features of specific rock types can effectively improve fine-grained classification performance. However, in a study by Bai and Yao [8] expanded to 15 lithologies (approximately 1000 images per class), validation accuracy plummeted to 63%, likely due to feature confusion effects caused by crossing mineral compositions in multi-class scenarios. Notably, Xiong and Liu achieved a validation accuracy of 76.31% in a study of 8,514 mesoscopic images of typical rocks (≥1,371 per class) from the main urban area of Chongqing [39]. Although identification accuracy showed an upward trend with increasing single-class sample sizes compared to the study by Bai and Yao [8], it remained significantly lower than that achieved by Zhang and Li [20], who utilized fewer samples per class (Table 2).

In applications involving the ResNet50 model (Table 2), Hu and Ye achieved 90% test accuracy based on 1,200 samples across 8 classes (~150 per class) [40], potentially due to the effective representation of mineral paragenetic associations by deep residual networks under limited categories. However, subsequent research presents contradictory trends: Wang and Liu [36], investigating 4 lithologies (596 per class), found that even after excluding broken rock masses with developed structural joints, model identification accuracy ranged only between 75% and 90%. Similarly, Ren and Zhang [44], in a study of over 60,000 samples covering more than 100 types, observed a significant drop in accuracy to just over 50%, despite increasing single-class samples to ~600. Furthermore, Zhang and Yi expanded single -class samples to 3,912 (27,384 total across 7 classes) [46], yet test accuracy reached only 74.1%.

The VGG-16 model exhibits similar patterns (Table 2). Yang and Xiong obtained 91.6% accuracy in a study of 221 cutting images across 5 classes (≥33 per class) [43]. However, when Dong and Zhang extended the research to 3,526 cutting images across 18 classes (up to 195 per class) [33], the test set lithology identification accuracy was 87.3%, again lower than the study by Yang and Xiong which used fewer samples per class [43]. Interestingly, Zhang and Tang [32], based on 6 classes with 300 samples each, reported a test set accuracy of 94.56%, reaffirming the trend where larger single-class sample sizes correlate with higher identification accuracy.

In summary, although some studies indicate that data augmentation can enhance model performance in specific scenarios, comprehensive analysis across single models like ResNet50, Inception-V3, and VGG-16, as well as cross-model synthesis (Figure 3), reveals that increasing the number of single-class rock images yields diminishing marginal returns on identification accuracy. When the number of categories exceeds the model's representational capacity, sample size growth may even lead to accuracy degradation. This contradictory phenomenon exposes the limitations of traditional data augmentation strategies: without synchronous optimization of model capacity and feature decoupling capability, merely increasing sample quantity is insufficient to breakthrough the theoretical bottlenecks of multi-class lithology identification**.**

Table 2 Performance Comparison of Convolutional Neural Networks in Rock Image Recognition Tasks

Model	Total images	Number of images per class	Test accuracy (%)	Reference	No.
Inception-V3	9 images (3 classes)	3	87~97%	[20]	1
Inception-V3	~15,000 images (15 classes)	~1000	63 %	[8]	2
Inception-V3	8,514 images (4 classes)	＞1371	76.31%	[39]	3
ResNet50	1,200 images (8 classes)	~150	~90%	[40]	4
ResNet50	2,384 images (4 classes)	596	75%~90%	[36]	5
ResNet50	60,000 images (>100 classes)	~600	50%	[44]	6
ResNet50	27,384 images (7 classes)	~3912	74.1%	[46]	7
VGG-16	221 images (5 classes)	35~64	91.6%	[43]	8
VGG-16	3,526 images (18 classes)	195	87.3%	[33]	9
VGG16	1,800 images (6 classes)	300	94.56%	[32]	10

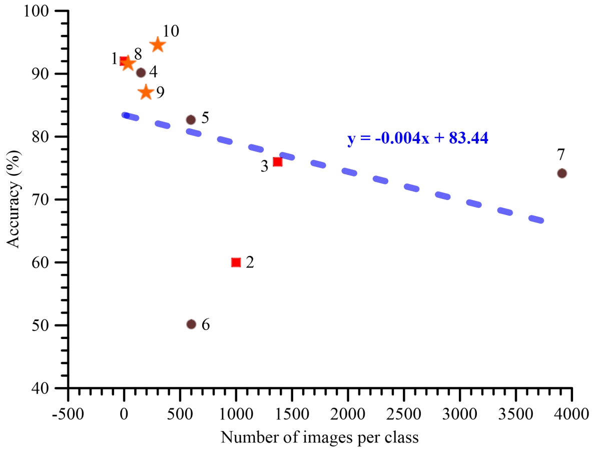


Figure 3 Correlation between Intelligent Rock Image Recognition Accuracy and sample size per class

Note: The blue dashed line represents the fitted trend for the aggregate data. Refer to Table 2 for data sources and corresponding serial numbers.

4.2 Resolution Enhancement Technologies: Capture and Limitations of Mesoscopic Features

In intelligent rock image recognition, the convergence of mineral compositions and the similarity of macroscopic characteristics constitute a dual challenge. For instance, Zhang and Tang found that models tend to misclassify shale as dolomite due to color similarity [32]. Alzubaidi and Mostaghimi discovered that ResNet-18 erroneously classified approximately 50 limestone images as sandstone and about 40 limestone images as shale [34]. Dong and Zhang noted that the identification accuracy for gray argillaceous siltstone was only 34.7% [33], with half being misclassified as dark gray silty mudstone. Tan and Tian observed that three types of tuff (RLCSF-Tuff, RSB-Tuff, RVMBV-Tuff) were frequently misjudged as the characteristically similar RCG-Tuff [10]. Bai and Yao found a 10% mutual misjudgment rate among dolomite [8], limestone, and marble due to close mineral compositions, and a 5%–10% misjudgment rate between gabbro and basalt.

To address these issues, considering that rocks with similar macroscopic features or mineral compositions often exhibit differences in mesoscopic features such as texture, roundness, and grain size [32, 38, 40], scholars have attempted to acquire high-precision rock image data to display these details, thereby improving model accuracy. For example, Xiong and Liu utilized a measuring electron microscope at 90x magnification to capture mesoscopic images of four typical rock samples—mudstone, sandy mudstone [39], argillaceous sandstone, and sandstone—from the main urban area of Chongqing, which showed obvious differences in color, flatness, grain prominence, and cementation forms (Figure 4 a-b). Building on this, a deep learning model for mesoscopic rock images was established using Inception-V3 and transfer learning. Results showed that sandstone identification accuracy in the validation set reached 97.28%. However, this study also illuminates the limitations of resolution enhancement: the identification accuracies for argillaceous sandstone and sandy mudstone, which have similar major components, were only 72.59% and 72.35%, respectively. Furthermore, argillaceous sandstone had a 14.02% probability of being mistaken for sandy mudstone, while sandy mudstone had a 12.18% probability of being mistaken for argillaceous sandstone, with mutual error rates exceeding 10%. Similarly, a recent study indicated that for limestone and tuff, which are visually close and difficult to distinguish with the naked eye (Figure 4 c-d), model differentiation remains problematic even with high-resolution images acquired via multiple magnifications and supplementary lighting [43], where the misidentification probability for limestone reached 11%.

Overall, for rock images with similar macroscopic features or mineral compositions, elevating image resolution does provide models with more learnable detailed features. However, judging from the aforementioned research results, substantial room for improvement remains in using resolution enhancement to increase identification accuracy, particularly compared to the accuracy achievable for rock images with significantly distinct macroscopic features.

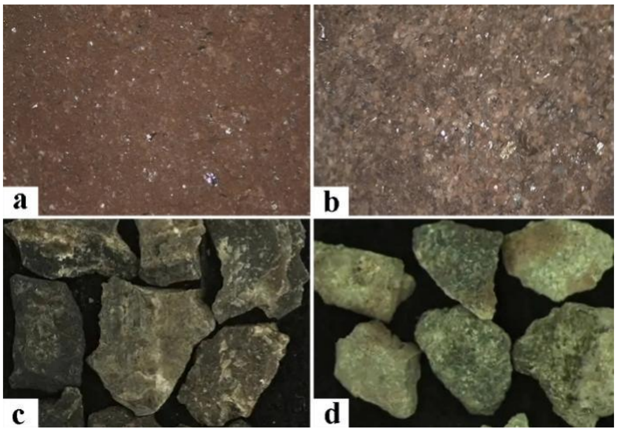


Figure 4 Examples of High-Resolution Rock Images: (a) Sandy mudstone and (b) argillaceous sandstone [39]; (c) tuff cuttings and (d) limestone cuttings [43]

4.3 Model Architecture Improvement: Adaptive Design for Complex Scenarios

Despite continuous optimization of rock image recognition models, the classification precision of existing methods still struggles to meet practical engineering requirements. Particularly in complex geological scenarios, model robustness and generalization capability face severe challenges [38, 44]. As shown in Figure 5, scholars have proposed various innovative improvement schemes tailored to different application scenarios.

In terms of feature extraction optimization, Zhang and Zhang introduced multi-scale dilated convolution attention blocks into ResNet-101 to address the need for rapid in-situ rock classification at tunnel faces (Figure 5a) [35], effectively enhancing fine-grained feature capture capabilities. To handle complex image processing demands, Zhang and Ye innovatively combined VGG16 with Principal Component Analysis (PCA) to construct a PCA-VGG16 model (Figure 5b) [2], significantly reducing computational complexity while ensuring accuracy. Feng and Gong built a Siamese convolutional neural network model based on AlexNet to balance global image information and local textural information of rock data [22]. Ran and Xue successfully resolved the issue of interference elements affecting classification results by constructing a deep CNN model [23]. Additionally, improved models can achieve substantial accuracy increases for images with similar macroscopic features. For instance, Liu and Wang adopted a simplified

VGG16 as the image feature extraction network within a Faster R-CNN deep learning object detection framework [11]; results showed that probability scores for visually similar basalts (surface amygdaloidal structures) and conglomerates (surface rounded shapes) exceeded 99%.

Regarding engineering application adaptation, Yang and He improved the AlexNet model (Figure 5c) [52], significantly enhancing identification accuracy for large-sized rock fragments during tunnel boring processes. Xu and Ma integrated a Region Proposal Network and detector within the Faster R-CNN framework [53], developing an intelligent lithology identification system suitable for on-site detection. Addressing the limitations of 2D images, Xu and Shi proposed an intelligent lithology identification method based on deep learning of rock images and elemental information [54], markedly improving identification accuracy for weathered rocks. Considering the typically small sample size of rock specimens, Zhang and Yi designed a new neural network model [46], MyNet, targeted at small-sample databases, which demonstrated accuracy superior to ResNet50 and VGG16 models with or without transfer learning (Figure 5d).

In the direction of model lightweighting, Tan and Tian improved the Xception model by introducing residual connection mechanisms [10], significantly reducing parameter count while maintaining performance. Xiao and Li integrated technologies such as ResNetv1d and deformable convolutions to improve Mask R-CNN (Figure 5e) [51], realizing real-time online detection of ore types. Yang and Xiong employed depthwise separable convolutions to improve the ResNet model (Figure 5f) [43], significantly increasing the classification efficiency of sedimentary rock cuttings.

A synthesis of existing research reveals that whether improving existing models (e.g., optimization of AlexNet/VGG16) or constructing new ones (e.g., multi-modal identification systems), the average accuracy of improved models on test sets has indeed achieved significant enhancement, despite potential concerns regarding generalization capabilities**.**

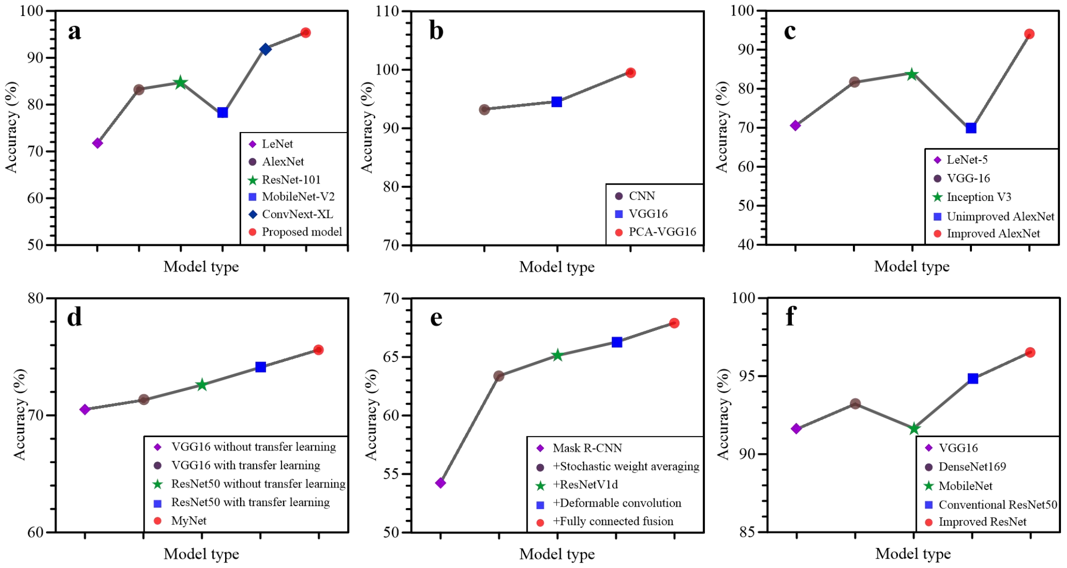


Figure 5 Performance Comparison between Optimized CNN Architectures and Benchmark Models for Rock Image Classification

Note: Data for panels (a)–(f) are derived from Zhang et al. [35], Zhang and Ye [2], Yang and He [52], Zhang and Yi [46], Xiao and Li [51], and Yang and Xiong [43], respectively.

5 CONCLUSION AND PERSPECTIVES

Research on intelligent rock image identification based on Convolutional Neural Networks has achieved significant progress, with CNNs demonstrating powerful feature extraction and classification capabilities that offer efficient solutions for this field. However, technical bottlenecks—including model selection, data scarcity, class imbalance, and environmental interference—continue to constrain model generalization capabilities and engineering applicability. While performance can be notably enhanced through data augmentation strategies, resolution enhancement technologies, and model architecture innovations, existing methods still possess limitations in complex lithology identification tasks. Therefore, constructing a robust intelligent rock image identification framework that overcomes the synergistic constraints of data, models, and the environment remains a shared challenge for both academia and industry.

Future development of intelligent rock image identification technology should focus on breakthrough research in the following dimensions:

(1) Construction of a multi-dimensional model selection assessment system. It is recommended to establish a three-dimensional decision model integrating task complexity, data distribution characteristics, and model structural parameters to form quantifiable model adaptation standards.

(2) Development of more efficient data augmentation strategies. This involves addressing data scarcity to enhance model generalization, while simultaneously fusing multi-modal data—such as rock images, elemental composition, and physical properties—to build a multi-dimensional feature representation system that improves discrimination of similar lithologies.

(3) Research on environmental adaptability technologies. This includes developing image enhancement and denoising algorithms, and designing robust models resistant to uneven illumination and weathering contamination, thereby mitigating the impact of environmental interference on image quality.

(4) Exploration of novel hybrid network architectures. Building upon the local feature extraction advantages of convolutional networks, research should introduce the global modeling capabilities of Transformers to construct hybrid architectures with multi-scale feature fusion**.**

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Liu X, Wang H, Jing H, et al. Research on intelligent identification of rock types based on Faster R-CNN method. IEEE Access, 2020, 8: 21804-21812.

[2] Zhang Y, Ye Y L, Guo D J, et al. PCA-VGG16 model for classification of rock types. Earth Science Informatics, 2024, 17(2): 1553-1567.

[3] Wu Lei, Zhai Xinwei, Wang Erteng, et al. Geochemical characteristics and formation environment of Jijitaizi ophiolite in Beishan area. Northwestern Geology, 2025, 58(1): 27-42.

[4] Meng Wuyi, Zhang Zhen, Gao Yongbao, et al. Mineral composition and geological significance of the newly discovered Wangzhuang gold deposit in Southern Qinling. Northwestern Geology, 2024, 57(4): 157-169.

[5] Liu Hao, Cui Junping, Jin Wei, et al. Geochemical characteristics and geological significance of granites in the eastern Songliao Basin. Northwestern Geology, 2024, 57(2): 46-58.

[6] Dai Xinyu, Zhou Bin, Li Xinlin, et al. Geochronology, geochemistry and tectonic significance of Miocene quartz monzonite intrusions in the north of Qitaidaban, West Kunlun. Northwestern Geology, 2024, 57(4): 191-205.

[7] Chen Yangyang, Duan Jun, Xu Gang, et al. Geochemical characteristics and tectonic significance of Late Triassic lamprophyres in Beishan area, Gansu Province. Northwestern Geology, 2024, 57(6): 78-94.

[8] Bai Lin, Yao Yu, Li Shuangtao, et al. Mineral composition analysis of rock images based on deep learning feature extraction. China Mining Magazine, 2018, 27(7): 178-182.

[9] Li Yan. Rock image recognition based on deep learning. Beijing: Beijing Forestry University, 2020.

[10] Tan Yongjian, Tian Miao, Xu Dexin, et al. Research on rock image classification and recognition based on Xception network. Geography and Geo-Information Science, 2022, 38(3): 17-22.

[11] Liu X, Wang H, Jing H, et al. Research on intelligent identification of rock types based on Faster R-CNN method. IEEE Access, 2020, 8: 21804-21812.

[12] Yuan Hang. Research on intelligent lithology identification method based on rock image feature learning. 2023.

[13] Marmo R, Amodio S, Tagllaferri R, et al. Textural identification of carbonate rocks by image processing and neural network: methodology proposal and examples. Computers & Geosciences, 2005, 31(5): 649-659.

[14] Chatterjee S. Vision-based rock-type classification of limestone using multi-class support vector machine. Applied Intelligence, 2013, 39: 14-27.

[15] Cheng Guojian, Yin Juanjuan. Rock thin section image classification based on SVM. Technology Innovation and Application, 2015, 5(1): 38.

[16] Izadi H, Sadri J, Bayati M. An intelligent system for mineral identification in thin sections based on a cascade approach. Computers & Geosciences, 2017, 99: 37-49.

[17] Chai H, Li N, Xiao C, et al. Automatic discrimination of sedimentary facies and lithologies in reef-bank reservoirs using borehole image logs. Applied Geophysics, 2009, 6: 17-29.

[18] Zhang F, Liu J, Lu X, et al. Spatial weighted graph-driven fault diagnosis of complex process industry considering technological process flow. Measurement Science and Technology, 2023, 34(12): 125143.

[19] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507.

[20] Zhang Ye, Li Mingchao, Han Shuai. Automatic identification and classification method of lithology based on deep learning of rock images. Acta Petrologica Sinica, 2018, 34(2): 333-342.

[21] Xu Zhenhao, Ma Wen, Lin Peng, et al. Intelligent lithology identification based on transfer learning of rock images. Journal of Basic Science and Engineering, 2021, 29(5): 1075-1092.

[22] Feng Yaxing, Gong Xi, Xu Yongyang, et al. Lithology identification method based on fresh rock surface images and Siamese convolutional neural networks. Geography and Geo-Information Science, 2019, 35(5): 89-94.

[23] Ran X, Xue L, Zhang Y, et al. Rock classification from field image patches analyzed using a deep convolutional neural network. Mathematics, 2019, 7: 755.

[24] Fan G, Chen F, Chen D, et al. Recognizing multiple types of rocks quickly and accurately based on lightweight CNNs model. IEEE Access, 2020, 8: 55269-55278.

[25] Wang C, Li Y, Fan G, et al. Quick recognition of rock images for mobile applications. Journal of Engineering Science and Technology Review, 2018, 11(4): 111-117.

[26] Fan G, Chen F, Chen D, et al. A deep learning model for quick and accurate rock recognition with smartphones. Mobile Information Systems, 2020, 2020: 1-14.

[27] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84-90.

[28] Chollet F. Xception: deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[29] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[30] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[31] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv, 2015.

[32] Zhang Z, Tang J, Fan B, et al. An intelligent lithology recognition system for continental shale by using digital coring images and convolutional neural networks. Geoenergy Science and Engineering, 2024, 239: 212909.

[33] Dong Wenhao, Zhang Huai. Lithology identification of cuttings based on transfer learning. Journal of University of Chinese Academy of Sciences, 2023, 40(6): 743-750.

[34] Alzubaidi F, Mostaghimi P, Swietojanski P, et al. Automated lithology classification from drill core images using convolutional neural networks. Journal of Petroleum Science and Engineering, 2021, 197: 107933.

[35] Zhang W, Zhang W, Zhang G, et al. Hard-rock tunnel lithology identification using multi-scale dilated convolutional attention network based on tunnel face images. Frontiers of Structural and Civil Engineering, 2024, 17(12): 1796-1812.

[36] Wang Xiaobing, Liu Lin, Wang Junqing, et al. Research on lithology identification of rock images based on convolutional neural network ResNet50 residual network. Geotechnical Engineering Technique, 2024, 38(3): 294-302.

[37] Xiong Feng, Liao Yifan, Cao Weiteng, et al. Automatic lithology identification based on convolutional neural network-deep transfer learning. Safety and Environmental Engineering, 2023, 30(4): 26-34.

[38] Ma Zedong, Ma Lei, Li Ke, et al. Multi-scale lithology identification based on deep learning of rock images. Bulletin of Geological Science and Technology, 2022, 41(6): 316-322.

[39] Xiong Yuehan, Liu Dongyan, Liu Dongsheng, et al. Automatic lithology classification method based on deep learning of mesoscopic images of rock samples. Journal of Jilin University (Earth Science Edition), 2021, 51(5): 1597-1604.

[40] Hu Qicheng, Ye Weimin, Wang Qiong, et al. Research on lithology identification based on big data of geological images. Journal of Engineering Geology, 2020, 28(6): 1433-1440.

[41] Chen Zhongliang, Yuan Feng, Li Xiaohui, et al. Interpretability study of deep transfer learning for images of plutonic intrusive rocks from Dabie Mountain. Geological Review, 2023, 69(6): 2263-2273.

[42] Xu Z, Ma W, Lin P, et al. Deep learning of rock microscopic images for intelligent lithology identification: neural network comparison and selection. Journal of Rock Mechanics and Geotechnical Engineering, 2022, 14(4): 1140-1152.

[43] Yang Lei, Xiong Chang, Liu Wenchao, et al. Research on lithology identification of cuttings based on improved ResNet deep residual network. Journal of Yangtze University (Natural Science Edition), 2023, 20(2): 11-19.

[44] Ren Wei, Zhang Sheng, Qiao Jihua, et al. Intelligent identification of rock and minerals based on deep learning. Geological Review, 2021, 67(S1): 1281-1282.

[45] Han Xinhao, He Yueshun, Chen Jie, et al. Research on intelligent identification of rock lithology based on Swin Transformer. Modern Electronics Technique, 2024, 47(7): 37-44.

[46] Zhang Chaoqun, Yi Yunheng, Zhou Wenjuan, et al. Small sample rock classification based on deep learning and data augmentation technology. Science Technology and Engineering, 2022, 22(33): 14786-14794.

[47] Liu Xiaobo, Wang Huaiyuan, Wang Liancheng. Faster R-CNN method for intelligent identification of rock types. Modern Mining, 2019, 35(5): 60-64.

[48] Xu Shuteng, Zhou Yongzhang. Experimental research on intelligent identification of microscopic ore minerals based on deep learning. Acta Petrologica Sinica, 2018, 34(11): 3244-3252.

[49] Theodoridis S. Machine learning: a Bayesian and optimization perspective. Academic Press, 2015.

[50] Bai Lin, Wei Xin, Liu Yu, et al. Rock thin section image recognition based on VGG model. Geological Bulletin of China, 2019, 38(12): 2053-2058.

[51] Xiao Chengyong, Li Qing, Li Hui, et al. Ore type detection algorithm based on improved Mask R-CNN. Sintering and Pelletizing, 2024, 49(2): 65-73, 106.

[52] Yang Z, He B N, Liu Y, et al. Classification of rock fragments produced by tunnel boring machine using convolutional neural networks. Automation in Construction, 2021, 125: 103612.

[53] Xu Z, Ma W, Lin P, et al. Deep learning of rock images for intelligent lithology identification. Computers & Geosciences, 2021, 154: 104799.

[54] Xu Z, Shi H, Lin P, et al. Intelligent on-site lithology identification based on deep learning of rock images and elemental data. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.