Asia Pacific Science Press

Journal of Advances in Engineering and Technology
ISSN(O):3005-7507
ISSN(P):3005-7493
Vol. 2 No.4 (2025)

Image Classification in Coal Production Using Deep Neural Networks: A Comprehensive Benchmarking Study

Wenmi Chai[1]*, Zhiyao Yang[2], Rui Zhao[3,4], Qian Xiang[5], Xinxin Niu[1], Ling Liang[1]

1.New Energy Technology Research Institute Co., Ltd., CHN ENERGY Investment Group Co., Ltd., Beijing, 102209, China

2.National Institute of Clean-and-Low-Carbon Energy, Beijing, 102209, China

3.Chengdu Jinjiang Center for Disease Control and Prevention, Chengdu, 610000, China

4.School of Economics and Management, Sichuan Normal University, Chengdu, 610101, China

5.Laboratory of Intelligent Control, PLA Rocket Force University of Engineering, Xi'an, 710025, China

Corresponding author: Wenmi Chai, wenmi.chai@ceic.com

Abstract: During the intelligent transformation of the coal industry, image classification technology plays a crucial role in process monitoring, quality inspection, and safety early warning. Taking the DsCGF-1 dataset in the coal production environment as the research object, this study conducts a multi-dimensional performance and efficiency evaluation on 12 mainstream deep learning models, aiming to establish industrial-level model selection criteria for intelligent coal separation. The results indicate that RepVGG-B3 exhibits the optimal comprehensive performance, with a test accuracy of 97.92%, a coal recall rate of 99.8%, and the best AUC value across all categories. Furthermore, RepViT-M3 achieves a test accuracy of 97.85% with a parameter count of merely 9.66M, demonstrating excellent lightweight characteristics, which makes it suitable for resource-constrained scenarios such as underground edge computing. This study establishes a model selection benchmark for coal separation, providing technical support for the development of intelligent sorting systems in industrial scenarios.

Keywords: Coal Separation; Image Classification; Deep Learning; Convolutional Neural Network; Model Benchmarking

Published: Dec 27, 2025

DOI: https://doi.org/10.62177/jaet.v2i4.958

1.Introduction

1.1 Research Background and Importance

As the cornerstone of China's energy security, coal plays an irreplaceable strategic role in ensuring the country's energy independence and self-reliance against the backdrop of the restructuring of the global industrial chain and energy supply chain[1]. In 2024, coal consumption accounted for 53.2% of China's total primary energy consumption[2]. In the advancement of the "carbon peaking and carbon neutrality" goals, the clean and efficient utilization of coal as well as its intelligent sorting directly influence the process of low-carbon transformation in the coal industry [1,2] . Traditional coal separation technologies struggle to meet the requirements of real-time performance and accuracy in modern industry [3,4]. With the development of computer technology and digital image acquisition equipment, machine vision technology, leveraging the advantage of non-contact detection, has become a core support for breaking through the efficiency bottleneck of traditional mine separation[4].

Early machine vision methods relied on the combination of manual feature extraction and machine learning-based classification. Numerous researchers have constructed ore recognition models using support vector machines (SVMs) [5–7], which exhibit favorable recognition performance on images collected in laboratory environments. However, these studies are detached from actual production scenarios and impose stringent requirements on both the acquisition environment and image resolution. The emergence of deep learning technology has offered a brand-new solution for coal image classification. A number of scholars [4,8–13] have adopted convolutional neural networks to replace the manual feature extraction process, which reduces the reliance on high-resolution images and enhances the classification accuracy of the models. Notably, Lv et al. [3] recently released the large-scale coal image dataset DsCGF, which includes coal images captured under both production and non-production conditions. This dataset provides a reliable benchmark for the systematic evaluation of the performance of deep learning models in real industrial scenarios.

1.2 Overview of Deep Learning Models

The aforementioned deep learning-based coal image classification research has laid the foundation for the technical advancement of intelligent sorting equipment in production environments. In recent years, deep learning model architectures have undergone continuous iterative optimization, forming a multi-technical route pattern that caters to diverse computational resource constraints and accuracy requirements, thereby providing abundant technical options for coal image classification tasks:

- ResNet [14]: Alleviates the gradient vanishing problem in deep networks through a residual connection mechanism, and simplifies training via residual mapping.
- Xception [15]: Adopts depthwise separable convolution to decompose the standard convolution process, significantly reducing computational complexity while ensuring performance.
- ResNeXt [16]: Introduces the concept of cardinality and parallel grouped convolution, enhancing feature diversity without increasing the network scale.
- MNASNet [17]: Pioneers platform-aware neural architecture search, with the latency of mobile devices as a constraint, achieving an accuracy of 75.2% with a latency of 78ms on Pixel phones.
- RepVGG [18]: Employs a reparameterization strategy of multi-branch training and single-path inference, featuring excellent model training accuracy and deployment inference speed.
- GhostNetV2 [19]: Proposes a hardware-friendly attention mechanism to enhance the extended features generated by low-cost operations, which can aggregate both local and long-range information simultaneously and optimize the performance of the network architecture.
- MobileOne [20]: Addresses the weak correlation between the traditional optimization objectives (such as FLOPs or parameter quantity) and the actual inference latency of devices by alleviating the architectural and optimization bottlenecks of existing efficient networks, realizing dual improvements in inference speed and accuracy on mobile devices.
- FastViT [21]: Optimizes memory access efficiency using the RepMixer module based on structural reparameterization, achieving a significant improvement in mobile inference speed while maintaining high accuracy, and demonstrating excellent latency-accuracy trade-off and cross-task generalization capabilities.
- ConvNeXt V2 [22]: Enhances the representation learning capability of pure convolutional networks by introducing a fully convolutional masked autoencoder framework and a global response normalization layer, achieving breakthrough performance in multiple visual tasks and providing a full range of efficient models with parameters ranging from 3.7M to 650M.
- RepViT [23]: Integrates the efficient design of Vision Transformer (ViT) into lightweight CNNs, achieving a latency of 1.0 millisecond with an accuracy exceeding 80% on the iPhone 12, and its inference speed increases by 10 times when combined with the Segment Anything Model (SAM).
- MobileNetV4 [24]: Adopts a universal Unified Inception Block (UIB) and a mobile-optimized attention mechanism to achieve comprehensive high efficiency across mobile hardware. Its large model reaches 87% accuracy on ImageNet with only 3.8ms inference latency on EdgeTPU.
- InceptionNeXt [25]: Decomposes large-kernel depthwise convolution along the channel dimension into four parallel branch-

es: small square kernels, two orthogonal strip kernels, and identity mapping. It maintains the advantage of large receptive field while significantly improving training throughput, and achieves a 0.2% accuracy improvement on ImageNet, serving as a lightweight architecture that balances performance and energy efficiency.

1.3 Research Motivation and Structure Arrangement

Each of the aforementioned models has its own advantages in terms of accuracy, efficiency, and deployment adaptability. However, most existing studies focus on the laboratory verification of a single model, lacking systematic evaluation of mainstream models in real production environments, which makes it difficult to form industrial-level model selection criteria. Based on this, this paper conducts a multi-dimensional evaluation on 12 advanced deep learning models, aiming to provide technical support for the research and development of intelligent coal separation equipment.

The structure of this paper is arranged as follows: Section 2 introduces the data and methods used in this study; Section 3 presents the multi-dimensional comparison results and corresponding analysis; Section 4 discusses the research limitations and future research directions; finally, the research achievements of the whole paper are summarized.

2.Material and Methods

2.1 Dataset

This study adopts the production-conditions subset of the Guobei Coal Preparation Plant in Anhui Province from the DsCGF-1 dataset released by Lv et al. [3] as the experimental data source. Collected from a real industrial environment, the dataset was acquired using an acA4096-40gc industrial camera manufactured by Basler, which was installed on the manual sorting conveyor belt. During the acquisition process, a stable illumination of 1200 (±100) Lux was maintained to ensure that the image quality meets industrial inspection standards. The dataset includes four categories: coal, gangue, unknown and foreign objects. All annotation work underwent strict quality control and was completed by an annotation team that passed professional assessments, ensuring the consistency and accuracy of the annotations.

The dataset was split into training, validation, and test sets in a 6:2:2 ratio based on the chronological order of image acquisition, effectively mitigating data leakage. As shown in Table 1, all images were uniformly preprocessed to a resolution of 224×224 pixels and normalized using the statistical parameters of the ImageNet dataset, providing a standardized data foundation for model training and evaluation.

Table 1: Subset division of the DsCGF-1 dataset and the corresponding sample sizes.

Dataset	Classes of Coal Images				Sum of Images
	Coal	Gangue	Unknown	Foreign Objects	
Training Set	2,926	57,799	2,201	1,890	64,816
Validation Set	975	19,266	734	630	21,605
Test Set	976	19,266	734	630	21,606
Total	4,877	96,331	3,669	3,150	108,027

This data construction method based on real industrial scenarios not only maintains the authenticity of the production environment but also ensures the reliability of model evaluation through scientific partitioning methods, laying a solid foundation for the performance evaluation of the deep learning models in this study.

2.2 Experimental Setup

The experimental environment of this study is built on the Windows 11 operating system. The hardware platform configuration includes an NVIDIA GeForce RTX 3090 GPU, an Intel Core i5-12400F processor (2.5 GHz, 6 cores and 12 threads), and 32 GB of physical memory. For the software environment, Python 3.12.11 is adopted as the programming language. The deep learning framework is based on PyTorch 2.8.0+cu129, and it is combined with the Torchvision 0.23.0+cu129 and Timm 1.0.19 libraries to implement model construction and training.

2.3 Data Preprocessing

The image preprocessing workflow adopts a standard computer vision task pipeline, with specific steps as follows:

1. Size Adjustment: All images are uniformly resized to a specified square size;

2. Format Conversion: Convert PIL images into PyTorch tensors, and normalize the pixel values to the range of [0, 1];

3. Standardization: Perform standardization using the statistical parameters of the ImageNet dataset, with a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225].

Regarding the configuration of the data loader, the key parameters are specified as follows: For the training set, the batch size is set to 128 with random data shuffling enabled, while the validation and test sets adopt a batch size of 256 without altering the data order, which ensures accurate evaluation of model performance. The optimizer used for all datasets is AdaBoB[26], with a total of 90 training epochs. The initial learning rate is set to 10^{-3}, and it is decayed by a factor of 0.1 every 30 training epochs. Additionally, a weight decay of 10^{-4} is applied to prevent overfitting.

2.4 Evaluation Metrics

This study employs standard evaluation metrics to quantitatively assess the model performance. For the four-class coal image classification task (coal, gangue, unknown and foreign objects), the evaluation is performed based on the confusion matrix, whose structure is presented in Table 2. This matrix intuitively illustrates the model's classification performance across different categories by conducting statistical analysis of the correspondence between true labels and predicted labels.

Table2: Four-Class Confusion Matrix

Class	Predicted coal	Predicted gangue	Predicted unknown	Predicted foreign object
Actual coal	TP_{11}	FP_{12}	FP_{13}	FP_{14}
Actual gangue	FP_{21}	TP_{22}	FP_{23}	FP_{24}
Actual unknown	FP_{31}	FP_{32}	TP_{33}	FP_{34}
Actual foreign object	FP_{41}	FP_{42}	FP_{43}	TP_{44}

Note: TP_{ii} denotes the number of correct predictions for class i, and FP_{ij} denotes the number of samples from class i incorrectly predicted as class j ($i \neq j$).

The classification performance metrics include Accuracy, Precision, Recall, and F1-Score. Accuracy reflects the proportion of all samples correctly classified; Precision denotes the proportion of correctly predicted samples among those detected as a specific class; Recall evaluates the proportion of actual positive samples of a target class that are correctly identified as such class; F1-Score provides a comprehensive measure of Precision and Recall. Based on the confusion matrix, the calculation formulas for each evaluation metric are as follows, where N denotes the total number of samples, represents the weight of class i.

Accuracy: Overall classification accuracy.

$$Accuracy = \frac{\sum_{i=1}^{4} TP_{ii}}{N} \tag{1}$$

Precision: Prediction accuracy calculated for each class i, followed by a weighted average.

$$Precision = \sum_{i=1}^{4} w_i \cdot Precision_i = \sum_{i=1}^{4} w_i \cdot \frac{TP_{ii}}{TP_{ii} + \sum_{j=1, j\neq i}^{4} FP_{ji}} \tag{2}$$

Recall: Recognition completeness calculated for each class i, followed by a weighted average.

$$Recall = \sum_{i=1}^{4} w_i \cdot Recall_i = \sum_{i=1}^{4} w_i \cdot \frac{TP_{ii}}{TP_{ii} + \sum_{j=1, j\neq i}^{4} FP_{ij}} \tag{3}$$

F1-Score: Harmonic mean of Precision and Recall, weighted by class.

$$F1\text{-}Score = \sum_{i=1}^{4} w_i \cdot 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{4}$$

Considering the class imbalance characteristics of the coal image dataset, all metrics are calculated using the weighted average method, where the weight of each class is the proportion of samples in that class.

3.Experimental Results and Analysis

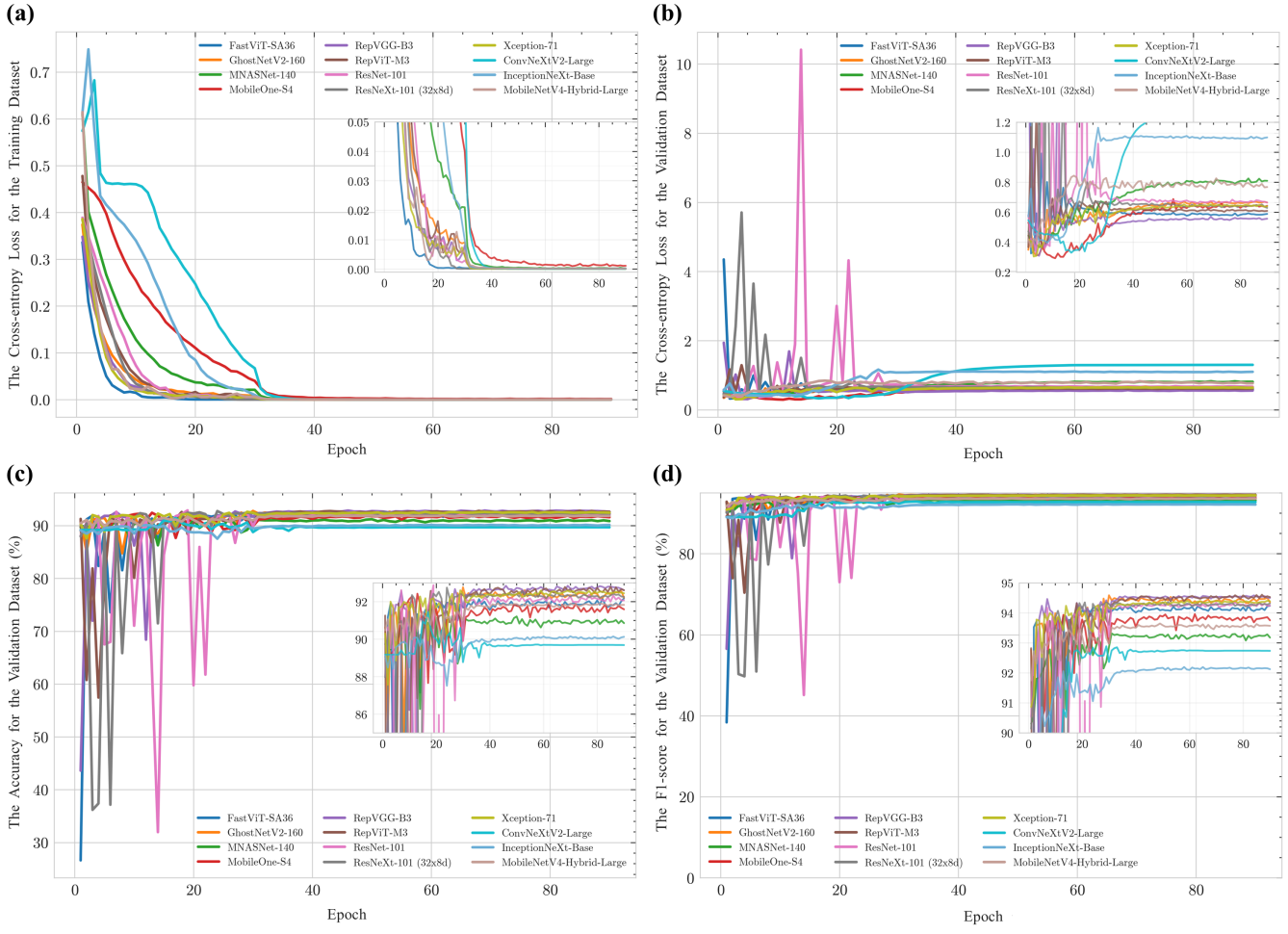
3.1 Model Training

By investigating the performance metrics of 12 deep learning models over 90 training epochs, this study systematically analyzed the dynamic evolution patterns of the models on both the training set and validation set. Figure 1 illustrates the

training processes of these 12 models on the training set and validation set. The loss values of all models on the training set exhibited a typical exponentially decaying convergence pattern, decreasing rapidly and then stabilizing (Figure 1a). Initially, the loss values ranged from 0.3 to 0.8; after 40 epochs, all converged to below 0.05, indicating that all models possessed the ability to effectively capture the essential features of coal images. However, significant differences were observed among different models in terms of convergence speed and training stability. Among them, FastViT-SA36 exhibited the fastest convergence speed: its loss dropped below 0.05 after 20 epochs, with minimal overall fluctuations. ConvNeXtV2-Large and InceptionNeXt-Base showed slower convergence and exhibited abnormal peaks in the early training stages, but stabilized after 40 epochs. From the locally enlarged view, it can be observed that MobileOne-S4 still presented slight fluctuations after 40 epochs, indicating its relatively weak training stability.

Figure 1b shows the loss of 12 models on the validation set, where a smaller loss indicates better generalization ability of the model. The loss of all models on the training set is less than 0.01, while their loss on the validation set stabilizes in the range of 0.3-0.8. Combined with the training set loss curves of each model, ConvNeXtV2-Large and InceptionNeXt-Base perform poorly: their convergence speed is lower than that of other models, and their validation loss on the validation set is higher than that of other models. Although ResNeXt-101 (32×8d) and ResNet-101 perform well on the training set, their validation loss fluctuates significantly in the first 25 epochs. After 40 epochs, the validation loss curves of all models gradually stabilize. The loss values of InceptionNeXt-Base, MNASNet-140, and MobileOne-S4 are greater than 0.7, which indicates that lightweight models do have an impact on performance while reducing model parameters.

Figure1: Training results of 12 deep learning models for coal image classification task on DsCGF-1 dataset. (a) Cross-entropy loss convergence curve on training dataset; (b) Cross-entropy loss generalization performance curve on validation dataset; (c) Accuracy curve on validation dataset; (d) F1-score curve on validation dataset.



Consistent conclusions can be drawn from the variation trends of accuracy and F1-score on the validation set (Figure 1c, d). ResNeXt-101 (32×8d) and ResNet-101 experienced severe fluctuations during the 0–20 epoch period: their minimum accuracy dropped below 40%, and the F1-score fell below 60%. This completely corresponds to the loss peaks of the two models on the validation set (Figure 1b), verifying the stability issues of these models. FastViT-SA36 also showed similar fluctuations but with a smaller amplitude. Eventually, the accuracy of each model also exhibited obvious differentiation: the accuracy of ConvNeXtV2-Large and InceptionNeXt-Base was around 90%, that of MNASNet-140 stabilized at approximately 91%, and the accuracy of most other models reached about 92%. A similar differentiated result was also observed in the F1-score, but the differentiation appeared earlier in the training epochs.

In terms of model performance effectiveness, Accuracy, as the core evaluation metric for classification tasks, reveals distinct hierarchical differences among the 12 models (Table 3). Models with high accuracy include ResNet-101 (92.88%), RepVGG-B3 (92.84%), RepViT-M3 (92.79%), GhostNetV2-160 (92.77%), and ResNeXt-101 (32×8d) (92.75%). These models exhibit superior performance in complex feature extraction and classification decision-making. Models with moderate accuracy cover MNASNet-140 (91.22%), ConvNeXtV2-Large (91.45%), MobileNetV4-Hybrid-Large (91.95%), and FastViT-SA36 (92.17%), whose performance is slightly inferior to that of the high-accuracy group. In contrast, InceptionNeXt-Base achieves the lowest validation accuracy (90.15%), showing poor overall classification performance. Models with high Precision generally have a Precision rate of over 99%, among which GhostNetV2-160 (99.96%), RepVGG-B3 (99.95%), and Xception-71 (99.94%) perform optimally, indicating an extremely low misjudgment rate in their prediction of positive samples. GhostNetV2-160 and RepViT-M3 achieve the highest F1-scores (both 94.59%), followed by RepVGG-B3 (94.54%) and Xception-71 (94.31%). This demonstrates that these models have greater advantages in balancing the classification performance of positive and negative samples, making them particularly suitable for task scenarios with imbalanced category distribution.

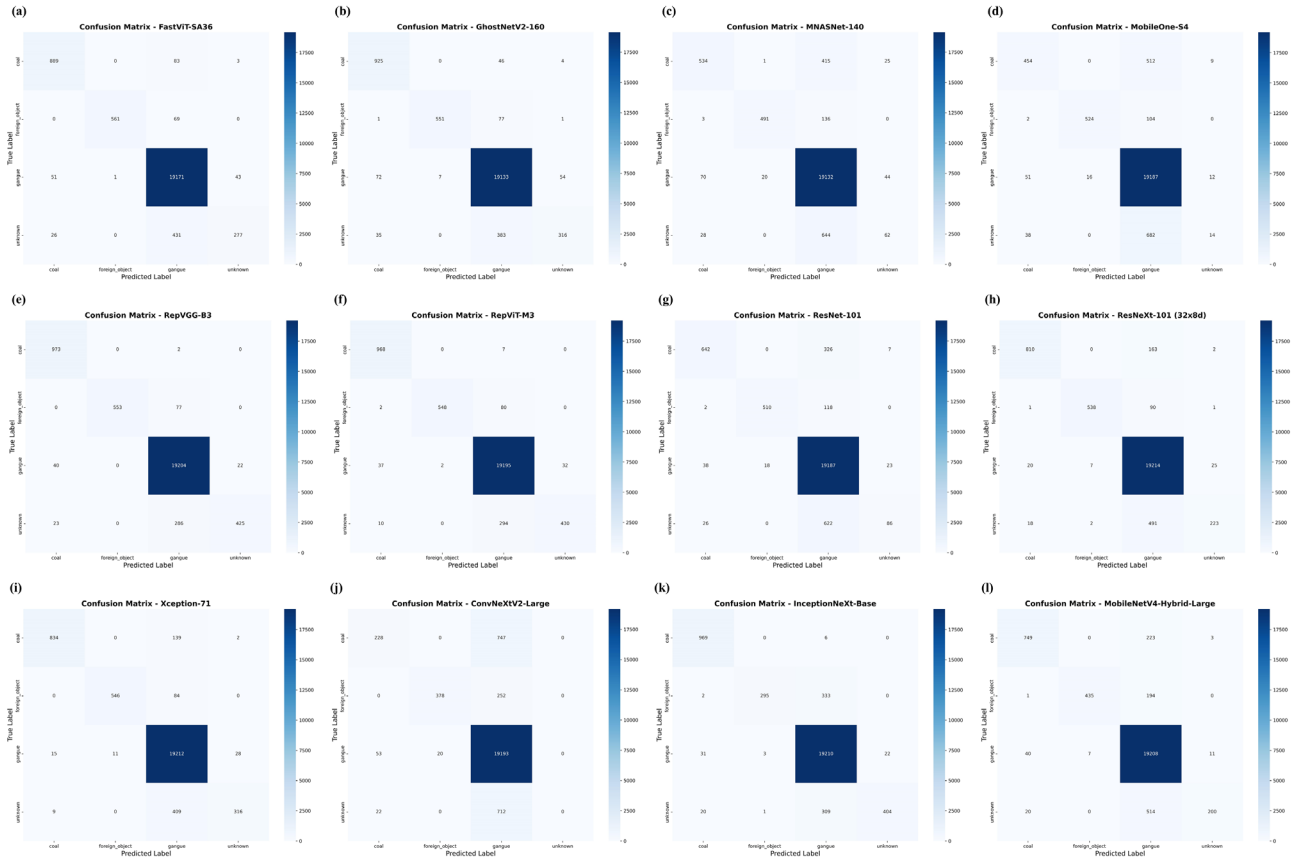
Table3: Results of 12 models on the DsCGF-1 validation set.

Model	Loss ($\times 10^{-2}$)	Accuracy (%)	Precision (%)	F1-score (%)
InceptionNeXt-Base	110.05	90.15	99.86	92.14
MNASNet-140	60.29	91.22	99.92	93.17
ConvNeXtV2-Large	32.98	91.45	96.48	92.69
MobileNetV4-Hybrid-Large	80.07	91.95	99.31	93.59
FastViT-SA36	58.88	92.17	99.94	94.24
MobileOne-S4	30.28	92.42	99.46	93.93
Xception-71	57.47	92.6	99.94	94.31
ResNeXt-101 (32x8d)	69.76	92.75	99.4	94.3
GhostNetV2-160	56.81	92.77	99.96	94.59
RepViT-M3	61.29	92.79	99.47	94.59
RepVGG-B3	56.93	92.84	99.95	94.54
ResNet-101	79.32	92.88	98.87	94.23

3.3 Test Results

To comprehensively evaluate the classification capability of each model in the coal image task, this study conducts an analysis of the performance of 12 models based on the results of the DsCGF-1 test set. Figure 2 presents the confusion matrix results of four image categories, and the models show significant differences in performance.

Figure 2: The confusion matrix of the 12 models on the DsCGF-1 test set: (a) FastViT-SA36, (b) GhostNetV2-160, (c) MNASNet-140, (d) MobileOne-S4, (e) RepVGG-B3, (f) RepViT-M3, (g) ResNet-101, (h) ResNeXt-101 (32x8d), (i) Xception-71, (j) ConvNeXtV2-Large, (k) InceptionNeXt-Base and (l) MobileNetV4-Hybrid-Large. X-axis: predicted labels; Y-axis: truth labels.



FastVit-SA36 achieved relatively high precision in classifying coal, gangue, and foreign objects, yet 432 unknown samples were misidentified as gangue (Figure 2a). GhostNetV2-160 exhibited similar performance to FastVit-SA36, with 383 unknown samples misclassified as gangue (Figure 2b). MNASNet-140 performed slightly worse: only 534 coal samples were correctly identified, while nearly 42% of coal samples, 21% of foreign objects, and 88% of unknown samples were misidentified as gangue (Figure 2c). MobileOne-S4 showed comparable performance to MNASNet-140, merely 454 coal samples were correctly recognized, and approximately 53% of coal samples, 16.5% of foreign objects, and 93% of unknown samples were misclassified as gangue (Figure 2d). RepVGG-B3 delivered the optimal performance, as nearly all coal samples were correctly identified. However, a small number of gangue and unknown samples were misclassified as coal. Additionally, the recall rate of this model for unknown samples—an area where other models performed poorly—reached approximately 58% (Figure 2e). RepViT-M3 exhibited similar performance to RepVGG-B3 but with slightly lower accuracy and precision across all categories (Figure 2f). ResNet-101 performed poorly, with a large number of coal, foreign object, and unknown samples misidentified as gangue (Figure 2g). ResNeXt-101 (32×8d) showed similar performance to ResNet-101 but was slightly more effective (Figure 2h). Xception-71 outperformed ResNet-101 yet was inferior to RepVGG-B3, as a considerable number of coal and unknown samples were still misclassified as gangue (Figure 2i). ConvNeXtV2-Large had very limited effectiveness, a large number of samples from other categories were misidentified as gangue. In particular, 77% of coal samples and 97% of unknown samples were misclassified as gangue, and none of the unknown samples were correctly recognized (Figure 2j). InceptionNeXt-Base fell into the second-tier performance group; its recall rate for coal classification reached 99.4%, but its performance in recognizing foreign objects and unknown samples was inferior to that of RepVGG-B3 (Figure 2k). MobileNetV4-Hybrid-Large showed similar performance to ConvNeXtV2-Large, with a large number of samples from other categories misidentified as gangue—more than 70% of foreign objects were misclassified as gangue

(Figure 2l).

By observing the confusion matrices, it can be found that more than 75% of the models perform poorly in classifying unknown samples and gangue. Some models, such as ConvNeXtV2-Large, even misidentify a large number of coal samples as gangue, which will undoubtedly cause economic losses in the production process. For coal production, image classification serves intelligent sorting, which requires as accurate identification of coal as possible to improve production and operation benefits. Therefore, the recall rate of the coal category is particularly important. Among the 12 models, RepVGG-B3 performs the best, with a recall rate of 99.8% and a precision rate of 93.9% for the coal category, which is of great significance to the coal production process.

Figure 3: The ROC curves of 12 models on the DsCGF-1 test set, with the abscissa representing the false positive rate and the ordinate representing the true positive rate. Among them, (a) coal, (b) gangue, (c) unknown and (d) foreign object.

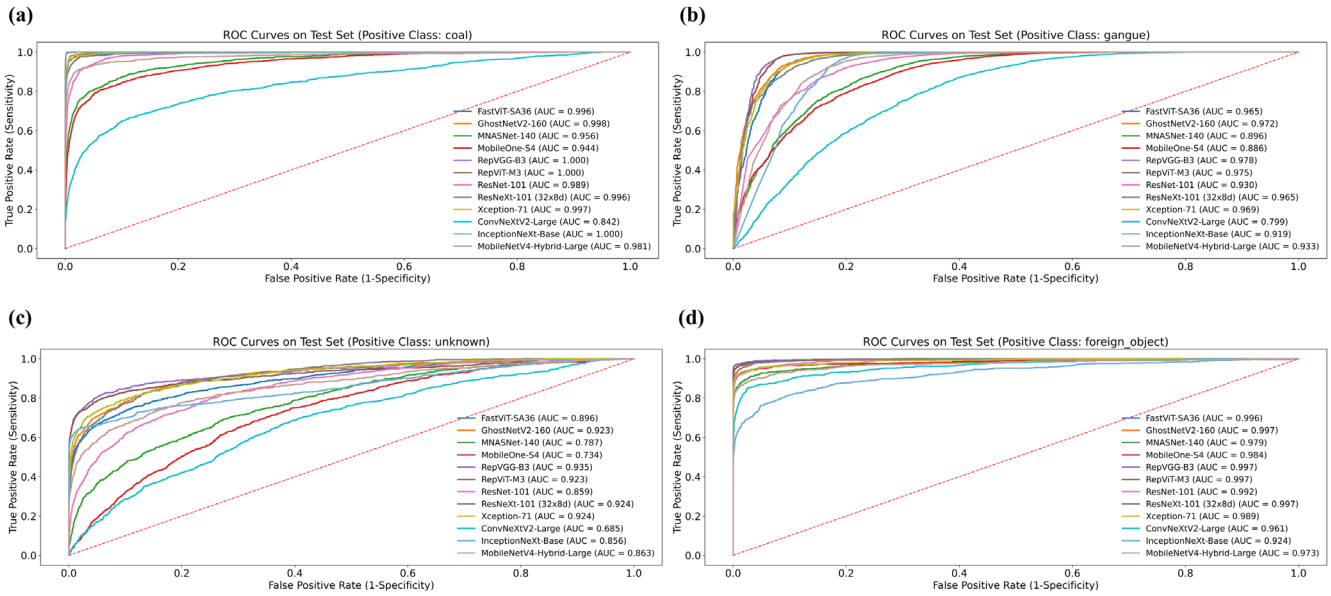


Figure 3 presents the ROC curve results of 12 models, where the AUC value (Area Under the ROC Curve) is used to evaluate the performance of classifiers. The closer the AUC value is to 1, the better the classifier performance; conversely, the closer the AUC value is to 0, the poorer the classifier performance. The AUC values of RepVGG-B3 for coal, gangue, unknown samples, and foreign objects are 1, 0.978, 0.935, and 0.997 respectively, which are the highest among the 12 models in all categories. This is consistent with the results of the confusion matrix. The AUC values of ConvNeXtV2-Large for the four categories are 0.842, 0.799, 0.685, and 0.961 respectively, ranking at a low level among the 12 models. Therefore, in terms of recognition performance alone, RepVGG-B3 is the optimal model on the DsCGF-1 test set, while ConvNeXtV2-Large has poor comprehensive performance and is not suitable as an intelligent coal separation model in this scenario. However, due to the limited equipment conditions in the coal production scenario, factors such as model training time, parameter quantity, and hardware requirements also need to be considered.

3.3 Model Complexity

Parameter scale and training time directly determine the deployment cost and training cost of a model. The results show (Table 4) that lightweight models such as MNASNet-140(5.84M), RepViT-M3(9.66M), GhostNetV2-16(11.12M), MobileOne-S4 (12.91M), and MobileNetV4-Hybrid-Large (36.49M) all have their parameter quantities controlled within 40M, which are quite consistent with the requirements of deployment scenarios with limited computing power, such as edge devices at coal mine wellheads. As models adapted to mobile lightweight scenarios, MNASNet-140 (5.84M), GhostNetV2-160, MobileOne-S4, RepViT-M3, and MobileNetV4-Hybrid-Large have small parameter quantities, which are much lower than those of large models like ConvNeXtV2-Large (196.43M) and RepVGG-B3 (120.53M).

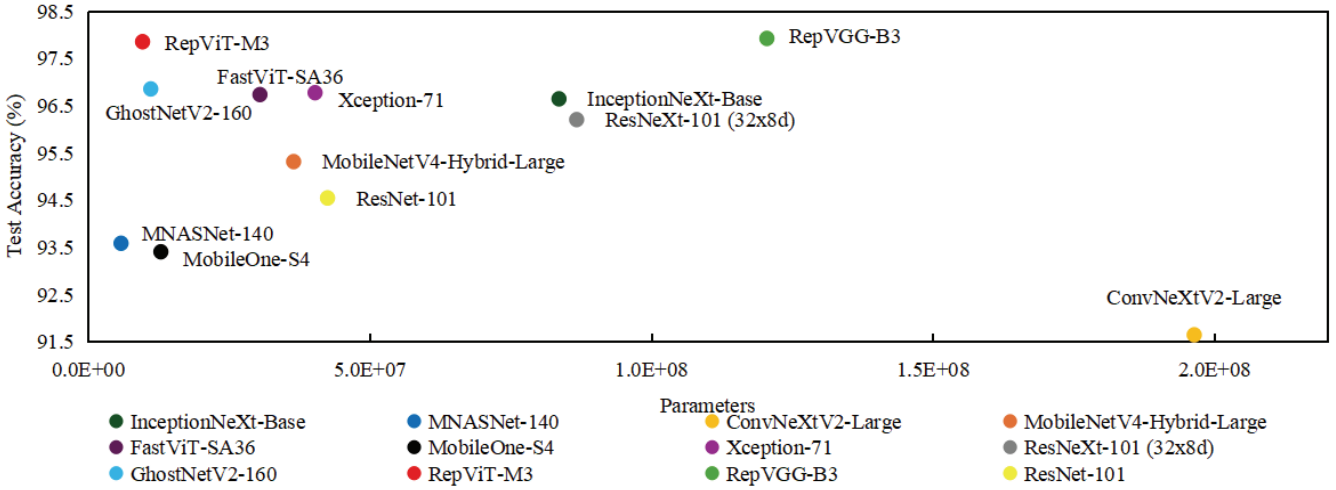
Table 4: Parameter quantity and training time of 12 models after 90 epochs on the DsCGF-1 set.

Model	Parameters(M)	Training Time (s)
InceptionNeXt-Base	83.61	20812.82
MNASNet-140	5.84	11774.49
ConvNeXtV2-Large	196.43	26651.59
MobileNetV4-Hybrid-Large	36.49	19378.11
FastViT-SA36	30.51	23373.23
MobileOne-S4	12.91	25883.84
Xception-71	40.3	29388.68
ResNeXt-101 (32x8d)	86.75	20791.29
GhostNetV2-160	11.12	28372.77
RepViT-M3	9.66	18442.13
RepVGG-B3	120.53	17857.85
ResNet-101	42.51	15773.18

In terms of training efficiency and time consumption, MNASNet-140 has obvious advantages among lightweight models, with a training time of only 11,774.49 seconds. As the fastest-trained model among the 12 models, it can reduce the model iteration cost. RepViT-M3 has a training time of 18,442.13 seconds, which is at a medium-low level. The training time of MobileNetV4-Hybrid-Large is 19,378.11 seconds, while that of MobileOne-S4 reaches 25,883.84 seconds. Although GhostNetV2-160 has only 11.12M parameters, its training time is 28,372.77 seconds. It is inferred that its attention mechanism increases the computational complexity of feature processing. Xception-71 has a medium-level parameter quantity, but its training time is the longest among all models.

Considering the parameter quantity and accuracy comprehensively (Figure 4), RepViT-M3 performs the best among lightweight models. This advantage benefits from the integration of ViT's efficient design into lightweight CNNs, enabling its pure convolutional structure to achieve Transformer-like global feature capture. The recall rate of coal classification reaches 99.3%, which is slightly lower than that of RepVGG-B3. In addition, compared with RepViT-M3, the parameter quantity of RepVGG-B3 is two orders of magnitude larger, but its training time is shorter. Moreover, RepVGG-B3 has distinct training and inference structures: during training, it adopts a multi-branch structure consisting of 3×3 convolution, 1×1 convolution and identity mapping to enhance feature extraction capability and alleviate gradient vanishing; during inference, through structural reparameterization, the multi-branch parameters are equivalently fused into a single 3×3 convolution layer, forming an extremely simple architecture similar to VGG, which further improves the inference speed.

Figure 4: Scatter plot of test accuracy and parameter quantity for 12 models, where the X-axis represents parameter quantity and the Y-axis represents test accuracy.



4.Conclusions and Recommendations

4.1 Conclusions

This study systematically evaluated the performance and efficiency of 12 deep learning models in the coal image classification task, and revealed the key patterns of the models in terms of training dynamics, classification accuracy, and resource consumption. The main findings are as follows:

1. In terms of the training and validation process, the training set loss of all models shows an exponential decay trend, converging to below 0.05 after 40 training epochs, which proves their ability to capture the core features of coal images. In the later training stage, there are obvious differences in convergence speed and stability: FastViT-SA36 converges the fastest (loss below 0.05 after 20 epochs), while ConvNeXtV2-Large and Inception Next-Base converge slowly with significant early fluctuations. The validation set loss stabilizes in the range of 0.3-0.8, and that of InceptionNext-Base, MNASNet-140 and MobileOne-S4 exceeds 0.7, indicating that lightweight models do affect performance while reducing parameters.

2. The core contradiction of model efficiency lies in the balance among parameter scale, training time, and performance. Lightweight models with parameters less than 40M, such as MNASNet-140 and RepViT-M3, are suitable for resource-constrained scenarios like edge devices at coal mine. However, training efficiency is not entirely determined by parameter quantity. MNASNet-140, with 5.84M parameters, achieves the shortest training time of 11,774.49 seconds, while GhostNetV2-160 (11.12M parameters) takes 28,372.77 seconds for training due to its attention mechanism. Among large models, RepVGG-B3 shows a advantage: although its parameter quantity (120.53M) is one order of magnitude higher than that of RepViT-M3, its training time (17,857.85s) is shorter, achieving a good training efficiency.

3. Considering classification performance, efficiency, and scenario adaptability, RepVGG-B3 and RepViT-M3 have emerged as the optimal representatives of large and lightweight models, respectively. RepVGG-B3 ranks first in test set accuracy (97.92%), coal recall rate (99.8%), and AUC values of various categories (coal AUC=1). The confusion matrix shows that it can almost correctly identify all coals, with an unknown object recall rate of 58%, making it the best model for intelligent sorting in coal production. RepViT-M3, with 9.66M parameters, achieves a high accuracy of 97.85% and a coal recall rate of 99.3%. Through its integrated architecture of “lightweight CNN + ViT design”, it achieves the optimal balance between performance and efficiency in resource-constrained scenarios. In addition, models have common shortcomings in category recognition: over 75% of models perform poorly in classifying unknown objects and gangue, and some models (such as ConvNeXtV2-Large) even misclassify 77% of coal as gangue, highlighting the importance of task adaptability in model selection.

4.2 Practical Significance and Industrial Application Suggestions

Based on the above conclusions, this study provides the following practical suggestions for the technological implementation of intelligent coal separation systems:

Differentiated model deployment schemes should be adopted for different application scenarios. Large-scale coal mine separation centers should give priority to RepVGG-B3, and leverage its advantage in high-precision coal identification to improve washing and separation efficiency, thereby reducing economic losses. For small and medium-sized coal mines or edge device, RepViT-M3 is recommended.

To promote the technological implementation and continuous optimization, it is also necessary to establish a guarantee mechanism from three dimensions: data, hardware, and mechanism. At the data level, it is suggested that coal mining enterprises cooperate with scientific research institutions to build an industry-level annotated dataset, which includes image samples from different regions, coal types, and working conditions, so as to solve the scenario limitations of the existing test set. At the hardware level, adaptive hardware solutions can be designed for the optimized models—for example, configuring a dedicated inference chip for RepVGG-B3 to improve speed, and developing edge computing modules for RepViT-M3 to reduce deployment costs. At the mechanism level, a closed-loop mechanism of “model performance monitoring - data update - fine-tuning optimization” can be established: misjudged samples from the separation site are collected regularly, and model parameters are optimized through incremental training to ensure that model performance is continuously adapted to changes in working conditions.

In addition, for enterprises with strong technical capabilities, model improvement can be carried out based on the conclusions of this study. For instance, drawing on the structural design idea of RepVGG-B3, customized models with "lightweight and reparameterization" can be developed to further compress parameters while retaining the advantage of high performance.

Future work can expand sample diversity, explore multi-modal data fusion, and develop customized models balancing lightweight and high performance, so as to promote the intelligent upgrading of coal separation technology.

Funding

This research was supported by the Science and Technology Project of CHN ENERGY Investment Group entitled "Research and Application of an Integrated Intelligent Simulation Platform for Full Industrial Chain Production and Operations Based on Digital Twin Technology." (Grant No. GJNY-23-176).

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Kang H, Xie H, Ren S, et al. (2022) Development Strategy of China's Coal Industry under the Reconstruction of Global Industrial Chain and Energy Supply Chain. Chinese Journal of Engineering Science 24: 26.

[2] National Bureau of Statistics of China. Statistical Communique on National Economic and Social Development of the People's Republic of China in 2024 [EB/OL]. (2025-02-28). https://www.stats.gov.cn/sj/zxfb/202502/t20250228_19588-17.html.

[3] Lv Z, Fan Y, Sha T, et al. (2025) A large-scale open image dataset for deep learning-enabled intelligent sorting and analyzing of raw coal. Sci Data 12: 403.

[4] Liu Y, Zhang Z, Liu X, et al. (2021) Deep Learning Based Mineral Image Classification Combined With Visual Attention Mechanism. IEEE Access 9: 98091–98109.

[5] Patel AK, Chatterjee S, Gorai AK (2019) Development of a machine vision system using the support vector machine regression (SVR) algorithm for the online prediction of iron ore grades. Earth Sci Inform 12: 197–210.

[6] Wang W, Lv Z, Lu H (2021) Research on methods to differentiate coal and gangue using image processing and a support vector machine. International Journal of Coal Preparation and Utilization 41: 603–616.

[7] Zhang L, Sui Y, Wang H, et al. (2022) Image feature extraction and recognition model construction of coal and gangue based on image processing technology. Sci Rep 12: 20983.

[8] Pu Y, Apel DB, Szmigiel A, et al. (2019) Image Recognition of Coal and Coal Gangue Using a Convolutional Neural Network and Transfer Learning. Energies 12: 1735.

[9] Si L, Xiong X, Wang Z, et al. (2020) A Deep Convolutional Neural Network Model for Intelligent Discrimination between Coal and Rocks in Coal Mining Face. Mathematical Problems in Engineering 2020: 1–12.

[10] Liu Q, Li J, Li Y, et al. (2021) Recognition Methods for Coal and Coal Gangue Based on Deep Learning. IEEE Access 9: 77599–77610.

[11] Liu H, Xu K (2023) Recognition of gangues from color images using convolutional neural networks with attention mechanism. Measurement 206: 112273.

[12] Cao Z, Fang L, Li R, et al. (2023) Research on image classification of coal and gangue based on a lightweight convolution neural network. Energy Science & Engineering 11: 3042–3054.

[13] Pengcheng Y, Heng Z, Xuyue K, et al. (2024) Lightweight detection method of coal gangue based on multispectral and improved YOLOv5s. International Journal of Coal Preparation and Utilization 44: 399–414.

[14] He K, Zhang X, Ren S, et al. (2016) Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, IEEE, 770–778.

[15] Chollet F (2017) Xception: Deep Learning with Depthwise Separable Convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, IEEE, 1800–1807.

[16] Xie S, Girshick R, Dollár P, et al. (2017) Aggregated Residual Transformations for Deep Neural Networks.

[17] Tan M, Chen B, Pang R, et al. (2019) MnasNet: Platform-Aware Neural Architecture Search for Mobile.

[18] Ding X, Zhang X, Ma N, et al. (2021) RepVGG: Making VGG-style ConvNets Great Again, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, IEEE, 13728–13737.

[19] Tang Y, Han K, Guo J, et al. (2022) GhostNetV2: Enhance Cheap Operation with Long-Range Attention.

[20] Vasu PKA, Gabriel J, Zhu J, et al. (2023) MobileOne: An Improved One millisecond Mobile Backbone.

[21] Vasu PKA, Gabriel J, Zhu J, et al. (2023) FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization.

[22] Woo S, Debnath S, Hu R, et al. (2023) ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders.

[23] Wang A, Chen H, Lin Z, et al. (2024) RepViT: Revisiting Mobile CNN From ViT Perspective.

[24] Qin D, Leichner C, Delakis M, et al. (2024) MobileNetV4 -- Universal Models for the Mobile Ecosystem.

[25] Yu W, Zhou P, Yan S, et al. (2025) InceptionNeXt: When Inception Meets ConvNeXt.

[26] Xiang Q, Wang X, Lei L, et al. (2025) Dynamic bound adaptive gradient methods with belief in observed gradients. Pattern Recognition 168: 111819.