frontiers | Frontiers in Energy Research

TYPE Original Research
PUBLISHED 13 January 2025
DOI 10.3389/fenrg.2024.1444877

OPEN ACCESS

EDITED BY
Yiliu Liu,
Norwegian University of Science and Technology, Norway

REVIEWED BY
Dora Luz Almanza-Ojeda,
University of Guanajuato, Mexico
Xinhua Liu,
China University of Mining and Technology, China

*CORRESPONDENCE
Tao Hu,
area001@yeah.net

RECEIVED 06 June 2024
ACCEPTED 26 December 2024
PUBLISHED 13 January 2025

CITATION
Hu T, Zhuang D and Qiu J (2025) An EfficientNetv2-based method for coal conveyor belt foreign object detection.
Front. Energy Res. 12:1444877.
doi: 10.3389/fenrg.2024.1444877

An EfficientNetv2-based method for coal conveyor belt foreign object detection

Tao Hu[1]*, Deyu Zhuang[1] and Jinbo Qiu[2]

[1]China Coal Technology and Engineering Group Shanghai Co., Ltd., Shanghai, China, [2]State Key Laboratory of Intelligent Coal Mining and Strata Control, Shanghai, China

The detection and recognition of foreign objects on coal conveyor belts play a crucial role in coal production. This article proposes a foreign object detection method for coal conveyor belts based on EfficientNetv2. Since MBConv and Fused-MBConv structures in EfficientNetv2 employ information compression and fusion strategies, which may lead to the loss of important information and affect the integrity of feature extraction, a hard shuffle attention (Hard-SA) mechanism is utilized to enhance the focus on important features and improve the representation ability of coal conveyor belts image features. To address the potential gradient disappearance issue during the backpropagation process of the network, an elastic exponential linear unit (EELU) activation function is introduced. Additionally, since the cross-entropy loss function may not be flexible enough to handle complex data distributions and may fail to fit the non-linear relationships between data well, a Polyloss function is adopted. Polyloss can better adapt to the complex data distribution and task requirements of coal mine images. The experimental results show that the proposed method achieves an accuracy of 93.02%, which is 2.39% higher than that of EfficientNetv2. It also outperforms some other state-of-the-art (SOTA) models and can effectively complete the detection of foreign objects on coal conveyor belts.

KEYWORDS

foreign object detection, EfficientNetv2, hard shuffle attention (Hard-SA), elastic exponential linear units (EELU), polyloss function

1 Introduction

In coal mine production, the coal conveyor belt plays a crucial role as the primary channel for coal transportation, directly impacting the efficiency of mining and coal transportation. Accurate detection and identification of foreign objects on coal conveyor belts have become essential tasks to ensure safe production in coal mines (Zeng et al., 2019). Common foreign objects on conveyor belts, such as bolts and large gangue, have the potential to cause scratches, tears, and coal stacking on high-speed running belts. Therefore, the timely detection and identification of foreign objects on coal conveyor belts are necessary for early warning and prompt handling of potential problems. Currently, different methods are employed for detecting foreign objects in coal conveyor belts, including manual identification, the x-ray method, and image processing (Einarsson et al., 2017; Zhang et al., 2021). Image processing methods encompass both object detection and image classification techniques (Zhang M. et al., 2022). Compared with object detection methods, the advantage of classification is that foreign objects can be

identified directly without pre-positioning. This provides a simpler process and enables a more efficient use of computational resources, allowing rapid identification.

The complexity of the mining environment poses challenges to existing image classification methods in detecting foreign objects in coal conveyor belts. Some researchers have explored the application of computer vision technology in the coal mining industry. Literature (Pu et al., 2019) employed VGG16 and transfer learning to recognize coal and gangue images, segregating coal and gangue. Since they used only 240 images, an accuracy rate of 82.5% was achieved. Literature (Dou et al., 2019) proposed the relief-support vector machine (relief-SVM) method to recognize coal and gangue based on image analysis, achieving accuracies of 92.57% and 92% on two datasets, respectively. Literature (Su et al., 2018) proposed an improved network based on LeNet-5, achieving a recognition rate of 95.88% for coal gangue. Literature (Hong et al., 2017) designed a gangue sorting system, con-structing a convolutional neural network (CNN) model based on AlexNet and transfer learning, and achieving an accuracy of 96.6%. Literature (Liu et al., 2021) combined deep learning and transfer learning to construct four CNN models with different depths and structures based on VGG16, VGG19, InceptionV3, and Res-Net50, with accuracy rates of 85.47%, 86.89%, 88.06%, and 90.91%, respectively. Literature (Hu et al., 2022) optimized the hyperparameters of CNN models using Bayesian algorithms for the quick and accurate identification of coal and gangue. Literature (Liu et al., 2023) proposed three different deep learning-based mineral image data augmentation models based on generative adversarial networks, addressing the issue of insufficient labeled images in the coal mining field. Literature (Zhang J. et al., 2022) proposed a novel algorithm that combines histogram equalization (HE) and Laplace algorithm. Then, the YOLOv5 model was used to identify the samples, with the recognition accuracy of this method for 50 common minerals reaching 95.6%. In the literature (Önal et al., 2020), AlexNet was used to classify minerals directly collected and photographed from the site, achieving an accuracy of 92.3% in an experiment that included a total of 1,491 images across 8 categories. Literature (De Lima et al., 2019) successfully classified microfossils, core images, rock micrographs, and hand sample images of rocks and minerals using a light-weight mobilenetv2 model, achieving an accuracy rate of 98%. Literature (Fan et al., 2020) established a rock image recognition model based on the lightweight network architecture ShuffleNet, combined with transfer learning methods. The recognition model achieved an accuracy of 97.65% on the PC testing dataset and 95.30% on the smartphone testing dataset.

Current research is primarily focused on the detection of coal gangue, while there is relatively less research on other foreign objects, such as anchor rods on coal conveyor belts. Due to the intricate nature of the underground environment in coal mines, challenges such as a substantial amount of dust, inconsistent humidity, poor lighting, and difficulties in collecting sample data arise. Current detection methods in the coal mining field still encounter issues of poor robustness, high model complexity, and extensive parameters. To address these challenges, this article proposes a foreign detection method for coal conveyor belts based on EfficientNetv2. The main contributions of this article are as follows:

(1) By integrating a hard-shuffle attention (Hard-SA) mechanism, the proposed method enhances the focus on important features, addressing the potential loss of image information caused by the MBConv and Fused-MBConv structures in EfficientNetv2.
(2) To mitigate the gradient disappearance issue during backpropagation, the elastic exponential linear unit (EELU) activation function is introduced. This enhancement ensures better network stability and improves the learning process, contributing to more accurate detections.
(3) The adoption of the Polyloss function addresses the limitations of the cross-entropy loss function in handling complex data distributions and fitting nonlinear relationships.

The organization of the remaining sections of the article is as follows: The second section introduces the materials and methods, detailing the dataset description, data augmentation techniques, and the proposed method. The third section covers the experiments and results, including the experimental environment and parameter settings, model evaluation indicators, experimental results and discussion. The fourth section presents the conclusions of the article.

2 Materials and methods

2.1 Dataset

The experimental data came from the mine dataset published by china university of mining and technology (https://github.com/CUMT-AIPR-Lab/CUMT-AIPR-Lab?tab=readme-ov-file). A total of 4,800 images of coal belt conveyor were collected, divided into three categories: large gangue, bolts, and normal coal. Each category includes 1,600 images, the training and test sets were split in a ratio of 8:2, which means 3,840 training images and 960 testing images.

2.2 Data augmentation

Although the data augmentation method using neural architecture search (NAS) for automatic search is effective, its limitation lies in the need to balance search efficiency and data augmentation performance. To address this issue, this article used the trivial augment (TA) (Müller and Hutter, 2021) data augmentation strategy. In contrast to previous data augmentation techniques, TA stands out for its simplicity and efficiency. Unlike other methods that involve complex parameter tuning and multiple data augmentation techniques per image, TA opts for a straightforward approach. Each image undergoes a random selection of an enhancement operation, followed by a random determination of its enhancement amplitude. This straightforward process ensures a streamlined and efficient image enhancement without the need for hyperparameter searches. Therefore, compared to auto-augment (AA) (Cubuk et al., 2018), population based augmentation (PBA) (Ho et al., 2019), and even rand augment (RA) (Cubuk et al., 2020), its search cost is almost negligible. After data augmentation, the amount of data is doubled.

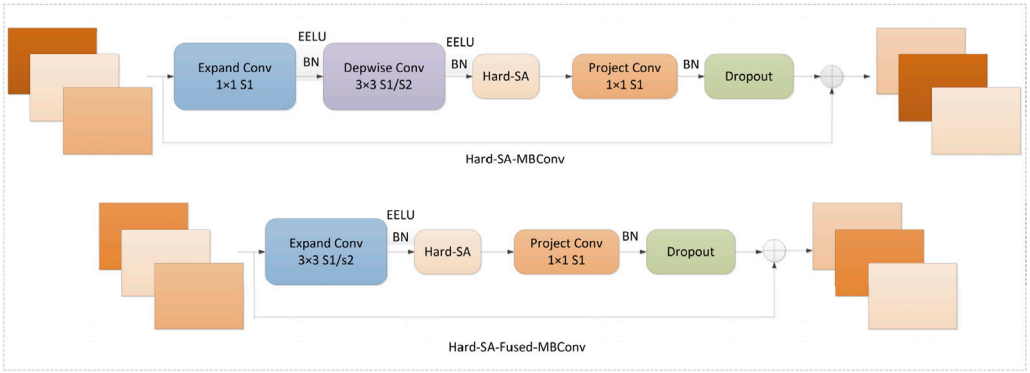

FIGURE 1
Hard-SA-MBConv and Hard-SA-Fused-MBConv structure.

2.3 Proposed method

EfficientNetv1 (Tan and Le, 2019) introduced a lightweight network structure that adjusts the resolution, depth, and width dimensions of the input image to enhance network performance. Building upon EfficientNetv1, EfficientNetv2 (Tan and Le, 2021) addressed the issue of slow speed caused by the use of deep convolution (DW) in the MBConv module of EfficientNetv1's shallow networks. EfficientNetv2 successfully reduced the number of parameters and complexity of model training by adaptively balancing these three dimensions, significantly improving model performance. The MBConv module in the EfficientNetv2 network model includes the squeeze-and-excitation (SE) attention mechanism, which solely focuses on encoding information between channels, overlooking the significance of positional information. This results in incomplete feature extraction of object through the attention mechanism, leading to lower classification accuracy in the model. Additionally, the derivative of the SiLU activation function may approach very small values for larger negative inputs, potentially causing gradient vanishing during back-propagation. This phenomenon can pose challenges in training deep neural networks. In the realm of deep learning classification networks, the cross-entropy loss function and the focus loss function are commonly used. However, a good loss function should ideally have a more flexible form and be tailored for different tasks and datasets. This is particularly important for coal mining data, where considerations for the underground lighting environment and various noise factors are crucial.

Motivated by EfficientNetv2, this article presents an improved network architecture. Initially, we incorporate Hard-SA module within MBConv and Fused-MBConv. This enhancement amplifies the net-work's capacity to prioritize crucial feature maps. Additionally, we utilize the EELU activation function, which addresses the potential issue of vanishing gradients during backpropagation. The utilization of EELU ensures a more stable and efficient training process. The refined structures, namely, Hard-SA-MBConv and Hard-SA-Fused-MBConv are as shown in Figure 1.

Additionally, a more flexible Polyloss function is adopted to improve training efficiency. This adaptive loss function enables the network to better adapt to the mining dataset, ultimately improving detection accuracy. Before training, the images undergo preprocessing to ensure consistency in formatting and normalization. The preprocessed data is then fed into the proposed network for training. The framework of the proposed method is shown in Figure 2.

2.3.1 Hard-SA mechanism

Assuming C, H, and W represent the number of image channels, height, and width, respectively. For feature map $Y \epsilon R^{C\times H\times W}$, the shuffle attention (SA) initially divides Y into D groups along the channel dimension, $Y = [Y_1, Y_2, \ldots Y_D]$, and $Y_k \epsilon R^{C/D\times H\times W}$. Each sub-feature Y_k gradually captures a specific semantic response during the training process (Zhang and Yang, 2021). Subsequently, generating the corresponding importance coefficient for each sub-feature through an attention module. At the beginning of each attention unit, the input of Y_k is split into two branches along the channels dimension. One branch generates a channel attention map, while the other branch produces a spatial attention map. To fully capture channel-wise dependencies, global averaging pooling (GAP) is employed. This process generates channel-wise statistics denoted as $u \in R^{C/2D\times 1\times 1}$, which can be calculated as Equation 1:

$$u = F_{gp} = \frac{1}{H \times W}\sum_{a=1}^{H}\sum_{b=1}^{W} Y_{k1}(a,b) \tag{1}$$

In addition, a compact feature is developed to facilitate accurate and adaptive selection. This is achieved through a straightforward gate mechanism implemented using a Hard-sigmoid activation function (Dou et al., 2023). Given an input x, the output of the Hard-Sigmoid function is defined as Equation 2:

$$f(x) = hard_{sig(x)} \begin{cases} = 0 & \text{if } x < -2.5 \\ = 0.2 * x + 0.5 & \text{if } -2.5 \le x \le 2.5 \\ = 1 & \text{if } x > 2.5 \end{cases} \tag{2}$$

The final output of channel attention can be obtained as Equation 3,

$$Y'_{k1} = \sigma(W_1 u + b_1).Y_{k1} \tag{3}$$

where, W_1 and b_1 are utilized to scale and shift u. Unlike channel attention, spatial attention is concerned with identifying the informative regions, complementing channel attention. The ultimate output of spatial attention is derived by Equation 4:

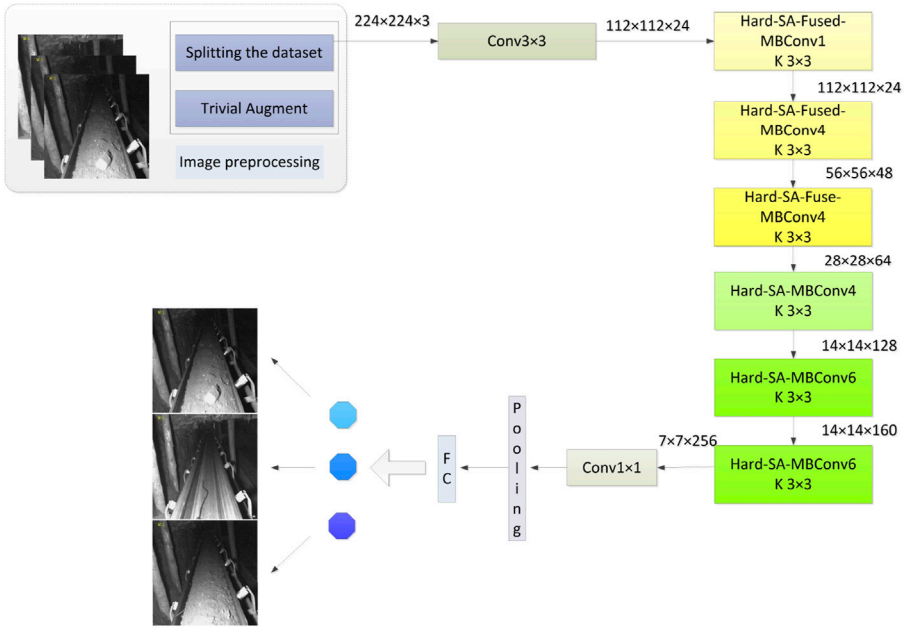

FIGURE 2
Framework of the proposed method.

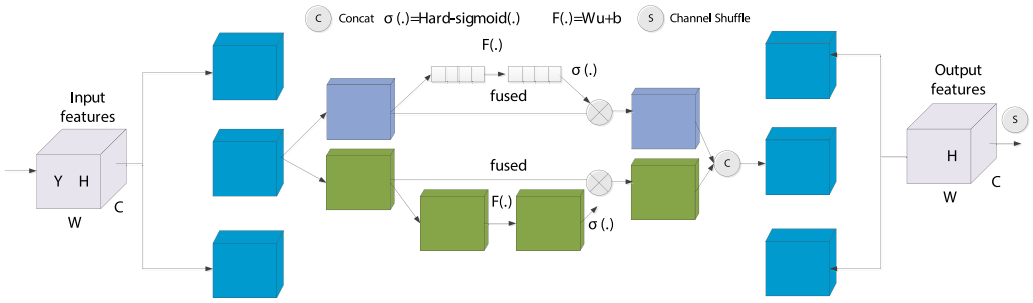

FIGURE 3
Hard-SA mechanism structure.

$$Y'_{k2} = \sigma(W_2 .GN(Y_{k2}) + b_2).Y_{k2} \tag{4}$$

where GN is utilized to obtain spatial-wise statistics. The role of W_2 and b_2 are the same to that of W_1 and b_1. Connecting the two branches as Equation 5:

$$Y_k = Y'_{k1} + Y'_{k2} \tag{5}$$

Subsequently, all sub-features are aggregated, and the final output of the Hard-SA module has the same size as Y, facilitating its seamless integration into deep learning network structures. The structure of Hard-SA is shown in Figure 3.

2.3.2 Activation and loss function design

The elastic exponential linear unit (EELU) is a versatile activation function that combines the benefits of both ReLU and ELU-type activation functions (Kim et al., 2020). EELU adjusts the positive slope to mitigate overfitting, similar to EReLU and RReLU (Jiang et al., 2018; Banerjee et al., 2020), while also preserving the negative signal to mitigate the bias shift effect, similar to ELU. Notably, the positive slope of EELU is determined using a Gaussian distribution with a randomized standard deviation, which differs from the approach of using a simple uniform distribution to determine the scale of random noise seen in EReLU and RReLU. This allows EELU to introduce randomness into the activation function, which can help improve the generalization performance of neural networks. It can be represented as Equations 6, 7:

$$f(y_{a,b}{}^{(c)}) \begin{cases} = t_{a,b}{}^{(c)} y_{a,b}{}^{(c)}, & \text{if } y_{a,b}{}^{(c)} > 0 \\ = \alpha^{(c)}\left(e^{\beta^{(c)}} y_{a,b}{}^{(c)} - 1\right), & \text{if } y_{a,b}{}^{(c)} \le 0 \end{cases} \tag{6}$$

$$t_{a,b}{}^{(c)} = \max(o, \min(q_{a,b}{}^{(c)}, 2)), q_{a,b}{}^{(c)} \sim N(1, \sigma)\sigma \sim U(0, \epsilon), \epsilon(0, 1) \tag{7}$$

where $y_{a,b}{}^{(c)}$ is the value of the cth channel at position (a, b), $t_{a,b}{}^{(c)}$ is a coefficient extracted from a Gaussian distribution with a random standard deviation and a fixed mean. $\alpha^{(c)}$ and $\beta^{(c)}$ are learning parameters greater than zero, determined from the training sample. ϵ is the maximum standard deviation of the Gaussian distribution. In the training process, the slope of the positive part undergoes adjustments through a Gaussian distribution. In the testing phase, the slope is substituted with the expectation derived from a Gaussian distribution.

When training deep neural networks for classification and segmentation problems, cross-entropy loss and focal loss are commonly used. However, a good loss function should have a flexible form and can be customized for different tasks and datasets. Polyloss allows for easy adjustment of the importance of different polynomial bases based on the target task and dataset (Leng et al., 2022). For images of coal mines, with characteristics such as complex backgrounds and poor lighting, the flexibility of the PolyLoss function provides convenience. The classification loss function is decomposed into a series of weighted polynomials by Taylor expansion, as shown in Equation (8):

$$L_{poly-1} = \sum_{j=1}^{\infty} 1/j(1 - P_t)^j + \varepsilon(1 - P_t) \tag{8}$$

where ε is the polynomial coefficient, P_t represents the probability of the target-label prediction.

3 Experiments and results

3.1 Experimental environment and parameter settings

The experiment was performed on a 64-bit Windows 10 system, using Python 3.8, PyCharm, and PyTorch 1.13.0. Hardware included an Intel Core i7-12700 CPU, 64GB RAM, and an Nvidia RTX 3090 24G GPU. The experimental setup involved resizing images to 224 × 224 pixels and utilizing a batch size of 32. Adam optimizer with a momentum of 0.9 and a weight decay of 0.0001 was employed for optimization. The initial learning rate was set at 0.0125, and a CosineAnnealingLrUpdater strategy with a minimum learning rate of 0.0001 was implemented. The experiment ran for 150 epochs, with 1,280 channels configured. Initialization of EELU was done with α set to 0.25 and β set to 1. Both learning parameters underwent weight decay, and the hyperparamet ϵwas fixed at 1.0. Additionally, the Polyloss function was configured with a polynomial coefficient ε of 2.

3.2 Models evaluation indicators

In the experiments, accuracy, precision, recall, F1-score and other metrics are used to evaluate the detection performance. The formulas of these evaluation metrics are given as Equations 9–12:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

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

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{12}$$

where TP represents true positive, FP represents false positive, FN represents false negative, TN represents true negative. In experiments, the precision-recall (P-R) curve is commonly used to evaluate the trade-off between precision and recall. The larger the area under the P-R curve (average precision, AP), the better the balance between precision and recall, indicating superior performance. The receiver operating characteristic (ROC) curve can also help assess the performance of a classification model. By observing the shape of the ROC curve and the area under the curve (AUC) value, one can intuitively understand the model's ability to classify different categories. A higher AUC value indicates better performance of the classifier.

3.3 Experimental results and discussion

The experimental results presented in Table 1 showcase the effectiveness of the proposed method in classifying different objects: bolt, gangue, and normal. For the bolt class, the method achieved a precision of 93.02%, recall of 91.56%, and an F1-score of 92.28%. The AP and AUC values are notably high at 97.47% and 98.58%, respectively, indicating excellent performance. The gangue class, while exhibiting slightly lower precision at 90.86%, has the highest recall at 96.25%, resulting in an F1-score of 93.47%. Its AP and AUC values are also high at 97.13% and 98.81%, respectively, reflecting robust detection capabilities. The normal class shows the highest precision at 95.42%, with a recall of 91.25% and an F1-score of 93.29%. The AP and AUC for this class are 97.67% and 98.33%, respectively. These results indicate that the proposed method performs consistently well across all classes, maintaining a strong balance between precision and recall, and achieving high accuracy in challenging detection environments. Figures 4, 5 respectively show the P-R curves and ROC curves of different objects.

We tested the model using different activation functions and attention mechanisms, with the cross-entropy loss function. The experimental results in Table 2 compare the performance of different models based on various activation functions and attention mechanisms, "Ev2" represents EfficientNetv2, "Acc" represents accuracy. The "Ev2-SE" model, utilizing the SiLU activation function and SE attention mechanism, achieves an accuracy of 90.63%. Similarly, the "Ev2-CA" model, employing SiLU activation and the coordinate attention (CA) mechanism (Hou et al., 2021), achieves an accuracy of 91.88%. Introducing the SA mechanism with SiLU activation in the "Ev2-Hard-SA-SiLU" model improves accuracy to 92.19%. Notably, the "Ev2-Hard-SA-EELU" model, incorporating the EELU activation function with the SA mechanism, demonstrates the highest accuracy of 92.92%. These findings suggest that the combination of specific activation functions and attention mechanisms significantly impacts model performance, with the "Ev2-Hard-SA-EELU" model outperforming

TABLE 1 The classification results of proposed method for each class.

Objects	Pr (%)	Re (%)	F1-score (%)	AP (%)	AUC(%)
Bolt	93.02	91.56	92.28	97.47	98.58
Gangue	90.86	96.25	93.47	97.13	98.81
Normal	95.42	91.25	93.29	97.67	98.33

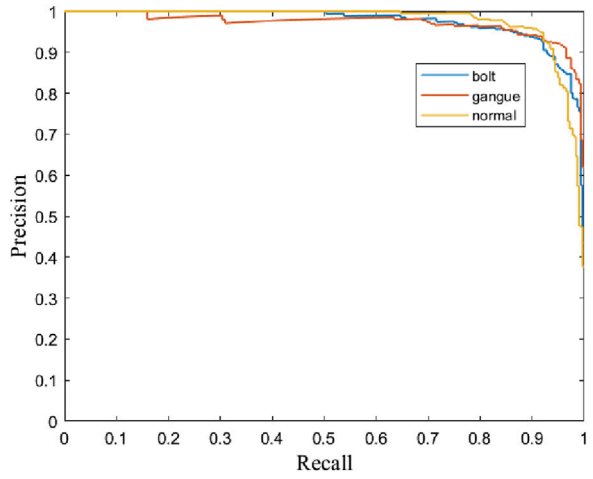

FIGURE 4
P-R curves of different objects.

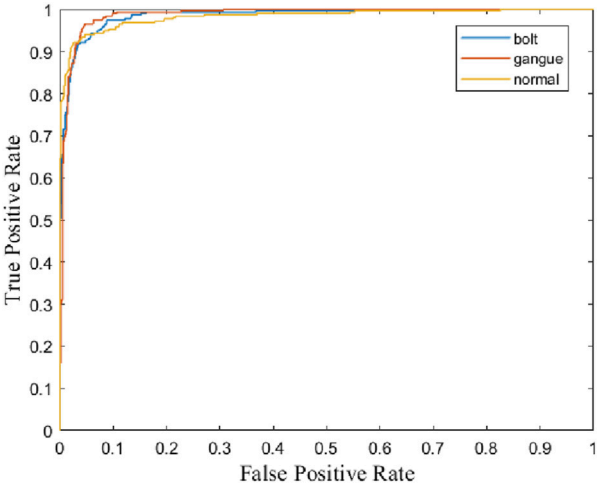

FIGURE 5
ROC curves of different objects.

others in terms of accuracy. Figure 6 shows the accuracy curves of different models.

To evaluate the impact of different modules, ablation experiments were performed. Table 3 presents the classification results of different improved modules using various combinations of enhancements on the EfficientNetv2 model. Each row represents a specific combination of techniques: Polyloss, EELU, and Hard-SA. The baseline EfficientNetv2 model without any enhancements achieves an accuracy of 90.63%. When the Polyloss module is added, the accuracy improves to 91.35%, indicating a notable enhancement in performance. Introducing the EELU module alone results in an accuracy of 91.98%, demonstrating a similar improvement. Adding the Hard-SA module on its own raises the accuracy further to 92.40%. Combining Polyloss with EELU yields an accuracy of 92.29%, while the combination of EELU and Hard-SA achieves 92.92%. Notably, combining Polyloss and Hard-SA without EELU results in an accuracy of 92.50%. The most significant improvement is observed when three modules Polyloss, EELU, and Hard-SA are combined, achieving the highest accuracy of 93.02%. This suggests that the combined effect of these three enhancements provides the most substantial performance boost, indicating their complementary strengths in improving detection accuracy. This method holds significant value for the detection of foreign objects in coal conveyor belts in dark well environments, as it offers improved accuracy and performance over the baseline model.

In Table 4, we compare the performance of several lightweight network models in detection tasks. These models include EdgeNext (Maaz et al., 2022), MobileNetv2 (Dong et al., 2020), EfficientNetv1 (Tan and Le, 2019), ShuffleNetv2 (Ma et al., 2018), and our proposed model. From the perspective of Acc, all models achieve relatively high performance, with our model slightly outperforming the others with an accuracy of 93.02%. This indicates that, despite having a slightly higher number of parameters (Params) compared to some other lightweight models, our model is able to achieve higher detection accuracy while maintaining model complexity. Regarding model complexity, MobileNetv2 and ShuffleNetv2 exhibit significant advantages in terms of smaller parameter counts and floating-point operations (Flops). However, this advantage comes with a sacrifice in accuracy. In contrast, EfficientNetv1 and EdgeNext strike a relatively good balance between accuracy and complexity. Notably, our model achieves the highest score in accuracy while maintaining a relatively low Flops of 0.6 G ($G = 10^9$). This further demonstrates the efficiency of our model in maintaining high performance while having a low computational cost. Figure 7 presents the confusion matrices for the comparison of these models, which clearly demonstrate the detection performance for each class.

Based on the provided confusion matrices for five different models, we can analyze the performance of each model on the three classes: 'bolt', 'gangue', and 'normal'. Our model demonstrates exceptional performance with minimal misclassifications, particularly in the gangue category where no instances were misclassified as normal. It can be seen that in the proposed method, 308 out of 320 gangue images were accurately identified, while only 293 were identified in the bolt images. The lowest recognition of normal images was only 292. MobileNetv2 also

TABLE 2 Comparative analysis of the performance of different models.

Model	Activation function	Attention mechanism	Acc (%)
Ev2-SE	SiLU	SE	90.63
Ev2-CA	SiLU	CA	91.88
Ev2-Hard-SA-SiLU	SiLU	SA	92.19
Ev2-Hard-SA-EELU	EELU	SA	92.92

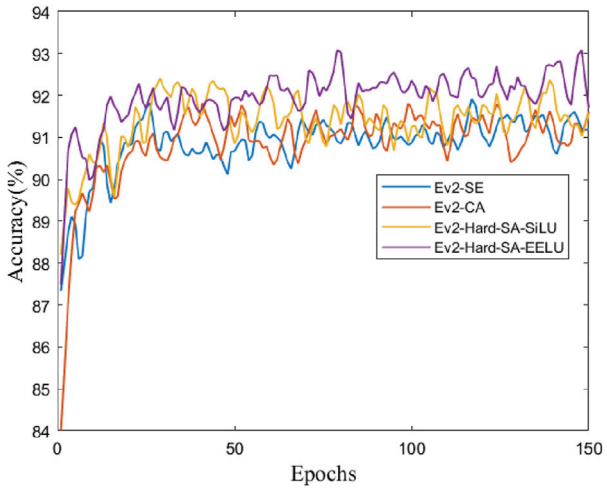

FIGURE 6
Comparison of test accuracy of the improved network model.

TABLE 3 Classification results of different improved module.

Method				Acc (%)
EfficientNetv2	Polyloss	EELU	Hard-SA	
√				90.63
√	√			91.35
√		√		91.98
√			√	92.40
√	√		√	92.29
√		√	√	92.92
√	√	√		92.50
√	√	√	√	93.02

shows strong results but with slightly higher misclassification rates, especially between normal and gangue. EdgeNext presents a higher rate of confusion, notably misclassifying bolts as normal and vice versa. ShuffleNetv2 performs well in classifying bolts but struggles significantly with the normal category, indicating a trade-off in its classification ability. EfficientNetv1 exhibits high accuracy in the gangue category but faces challenges in distinguishing between bolt and normal. Overall, our model achieves the best balance with the fewest misclassifications, highlighting its robustness and reliability in complex classification tasks, thereby underscoring its suitability for practical applications in the mining environment. Figure 8 visually presents the heatmaps generated by the proposed method and other models using Grad-CAM. As seen from Figure 8, the proposed method demonstrated superior capability in emphasizing the prominent features within the image.

TABLE 4 Comparison of lightweight network detection results.

Model	Acc (%)	Params (M)	Flops (G)
EdgeNext	90.83	5.28	0.96
MobileNetv2	92.29	2.23	0.32
EfficientNetv1	91.25	4.01	0.02
ShuffleNetv2	90.94	1.26	0.15
Ours	93.02	5.86	0.60

In Table 5, we compare the performance of several mainstream classification network models in detection tasks, including ResNet34 (Koonce and Koonce, 2021), Swinv2 (Liu et al., 2022), DeiT III (Touvron et al., 2022), MViTv2 (Li et al., 2022), and methods used for the classification of foreign matter in belts (Kou et al., 2023; Liu et al., 2024). From the perspective of Acc, all models achieve relatively high scores, but our model stands out with an accuracy of 93.02%. In terms of model complexity, our model significantly outperforms all other models. Its params is 5.86 M ($M = 10^6$), significantly lower than that of other models such as ResNet34's 21.29 M and Swinv2's 27.52M. Similarly, in terms of Flops, our model only requires 0.6 G, which is far lower than other models such as ResNet34's 3.68 G and Swinv2's 4.36 G. This advantage suggests that our model achieves high performance while having a lower computational cost, making it more suitable for resource constrained environments. Moreover, compared to other models, our model achieves a better balance between accuracy and complexity. Although models such as Swinv2 approach our model in accuracy, they fall far behind in terms of model complexity. The methods by (Kou et al., 2023; Liu et al., 2024), demonstrate relatively high accuracy, achieving 88.30% and 91.20%, respectively. However, both models exhibit higher computational demands compared to ours. Kou et al.'s method has 18.40 M parameters and requires 2.69 G, while Liu et al.'s model has 29.15 M parameters with 6.28 G. This indicates that our model has made significant progress in optimizing network structure, reducing redundant parameters, and lowering computational costs.

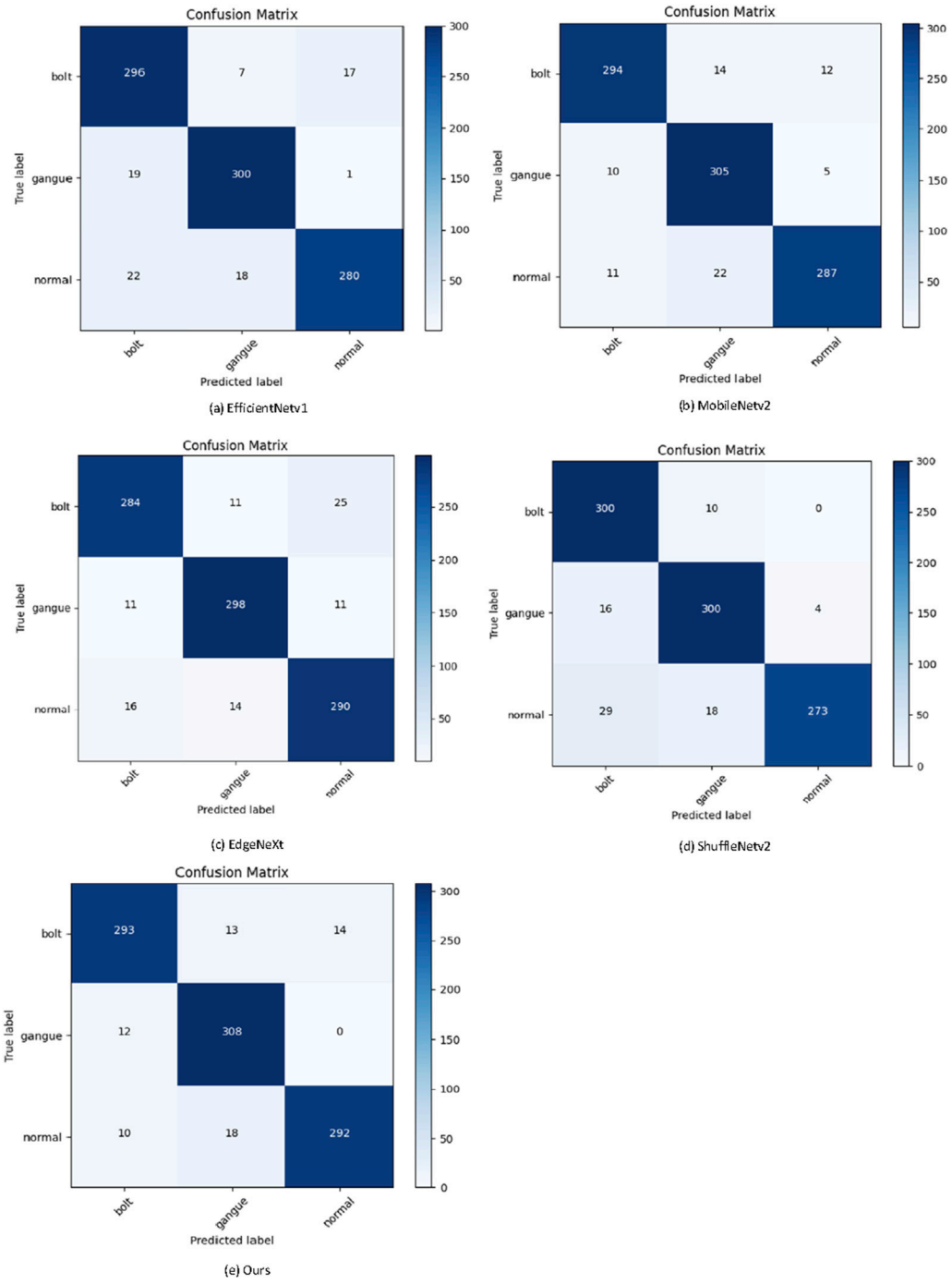

FIGURE 7
Confusion matrices of different models. **(A)** EfficientNetv1. **(B)** MobileNetv2. **(C)** EdgeNeXt. **(D)** ShuffleNetv2. **(E)** Ours.

Our model exhibits excellent performance in classification network detection tasks, achieving the highest level of accuracy while maintaining relatively low model complexity. This result provides strong support for the deployment of high-performance classification networks in practical applications, especially in resource-constrained scenarios.

The improved network architecture addresses the shortcomings of EfficientNetv2, enhancing the MBConv and Fused-MBConv modules while addressing vanishing gradient issues and providing a more flexible loss function for training. These modifications aim to improve the classification accuracy of foreign objects on coal belts and pave the way for more effective deep learning applications in this domain. The experimental results showed that the proposed method can not only effectively capture the prominent features of underground coal mine images, but also achieve the best performance in classification. Additionally, the smaller params and Flops make it easier to deploy the proposed model to edge devices.

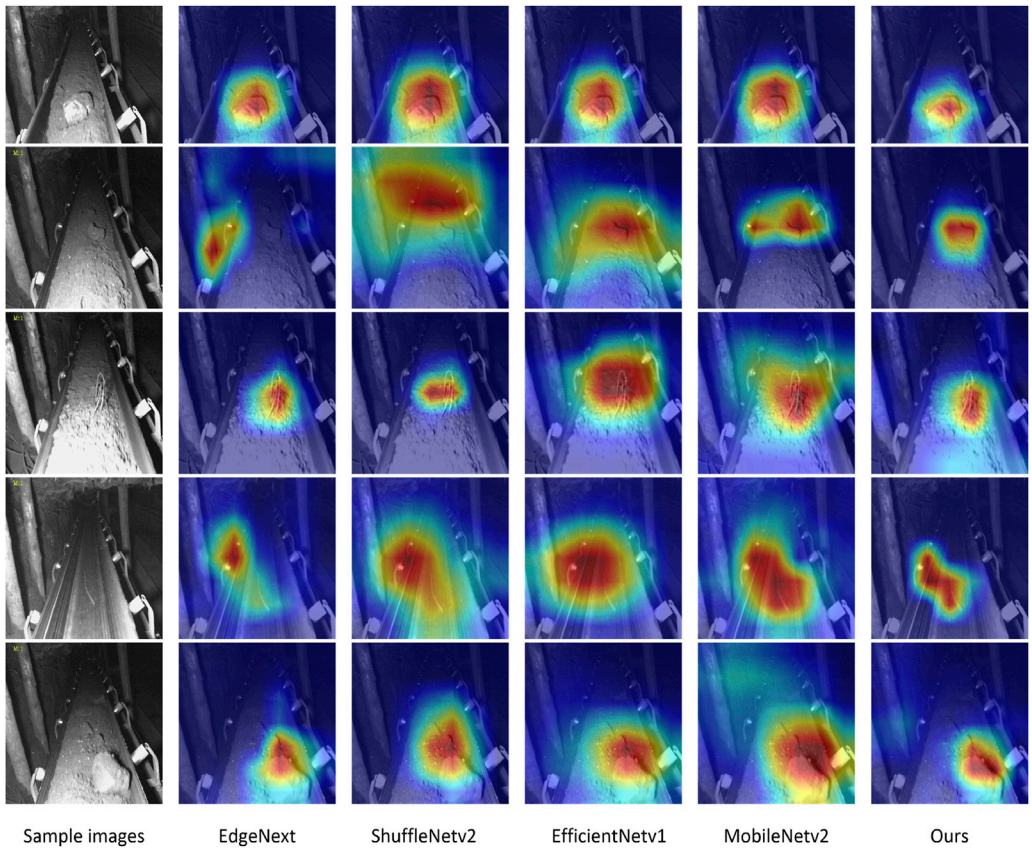

FIGURE 8
Comparison of heatmaps for different methods.

TABLE 5 Comparison of detection results for mainstream classification networks.

Model	Acc (%)	Params (M)	Flops (G)
ResNet34	90.31	21.29	3.68
Swinv2	92.19	27.52	4.36
DeiT III	90.73	21.80	4.24
MViTv2	91.88	23.41	3.99
Kou et al	88.30	18.40	2.69
Liu et al	91.20	29.15	6.28
Ours	93.02	5.86	0.60

Our proposed EfficientNetv2-based classification method is primarily designed for single-object classification, making it suitable for scenarios with one foreign body. To address cases with multiple foreign bodies, we suggest using sliding window or region-based approaches. This involves dividing the input image into overlapping patches or employing a sliding window technique for independent classification, enabling the detection of multiple foreign bodies in different regions of the belt. Additionally, if multiple foreign bodies are frequently encountered, we can enhance our pipeline with a YOLO-based object detection module for accurate localization, while EfficientNetv2-based ensures high classification accuracy for each detected object.

4 Conclusion

This article presents a method utilizing EfficientNetv2-based for detecting foreign objects on coal conveyor belts. To address the challenge of limited data, a simple data augmentation technique, TA was utilized during preprocessing to enhance the model's fitting ability. The proposed method integrates Hard-SA into the MBConv and Fused-MBConv modules of the EfficientNetv2 architecture and employs the EELU activation function within these modules. Additionally, a more flexible loss function Polyloss is utilized in the model. Experimental results demonstrate that this approach significantly improves detection performance. To evaluate the model's effectiveness, a quantitative analysis was conducted, comparing the proposed method's detection results with those of existing advanced models. Our method achieved a detection accuracy of 93.02% on datasets of coal conveyor belts. These findings indicate that the proposed neural network structure effectively learns the characteristics of foreign objects on coal conveyor belts, accurately classifying and recognizing these objects in coal mines. This study lays a solid foundation for the detection and recognition of foreign objects on coal conveyor belts in underground coal mines.

In future research, there are still some issues to explore:

(1) The dataset used in this study primarily originates from a publicly available dataset for foreign object detection on coal conveyor belts in an underground mine. While it includes three categories: bolts, gangue, and normal, the dataset is limited in both quantity and diversity of classes. To enhance the algorithm's generalization ability, it is advisable to supplement the dataset with additional data on foreign objects found on conveyor belts used for coal transportation, thereby increasing dataset diversity.
(2) Conveyor belts used for coal transportation are often subjected to harsh underground environments with low light conditions. Additionally, there is a considerable presence of dust and water mist. The robustness of the algorithm is crucial under such conditions. However, the study did not evaluate the improved algorithm's performance under different conditions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

TH: Funding acquisition, Methodology, Software, Visualization, Writing–original draft. DZ: Conceptualization, Funding acquisition, Supervision, Writing–review and editing. JQ: Project administration, Validation, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This article was supported by the National Development and Reform Commission Project (grant number: 0732118), and the Science Foundation of China Coal Technology and Engineering Group Shanghai Co., Ltd. (grant number: 02062222823Q).

Conflict of interest

Authors TH and DZ were employed by China Coal Technology and Engineering Group Shanghai Co., Ltd.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from Science Foundation of China Coal Technology and Engineering Group Shanghai Co., Ltd. The funder had the following involvement in the study: The funder approved the decision to submit the manuscript for publication.

Publisher's note

References

Banerjee, C., Mukherjee, T., and Pasiliao, E. (2020). Feature representations using the reflected rectified linear unit (RReLU) activation. *Big Data Min. Anal.* 3 (2), 102–120. doi:10.26599/BDMA.2019.9020024

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2018). Autoaugment: learning augmentation policies from data. Available at: https://arxiv.org/abs/1805.09501 (accessed on April 11, 2019).

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). "Randaugment: practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

De Lima, R. P., Bonar, A., Coronado, D. D., Marfurt, K., and Nicholson, C. (2019). Deep convolutional neural networks as a geological image classification tool. *Sediment. Rec.* 17 (2), 4–9. doi:10.2110/sedred.2019.2.4

Dong, K., Zhou, C., Ruan, Y., and Li, Y. (2020). "MobileNetV2 model for image classification," in *2020 2nd international conference on information technology and computer application (ITCA) (IEEE)*, 476–480. doi:10.1109/ITCA52113.2020.00106

Dou, D., Wu, W., Yang, J., and Zhang, Y. (2019). Classification of coal and gangue under multiple surface conditions via machine vision and relief-SVM. *Powder Technol.* 356, 1024–1028. doi:10.1016/j.powtec.2019.09.007

Dou, G., Zhao, K., Guo, M. E. I., and Mou, J. U. N. (2023). Memristor-based LSTM network for text classification. *Fractals* 31 (06), 2340040. doi:10.1142/s0218348x23400406

Einarsson, G., Jensen, J. N., Paulsen, R. R., Einarsdottir, H., Ersbøll, B. K., Dahl, A. B., et al. (2017). Foreign object detection in multispectral x-ray images of food items using sparse discriminant analysis. *Proceedings* 20, 350–361. doi:10.1007/978-3-319-59126-1_29

Fan, G., Chen, F., Chen, D., Li, Y., and Dong, Y. (2020). A deep learning model for quick and accurate rock recognition with smartphones. *Mob. Inf. Syst.* 2020, 1–14. doi:10.1155/2020/7462524

Ho, D., Liang, E., Chen, X., Stoica, I., and Abbeel, P. (2019). "Population based augmentation: efficient learning of augmentation policy schedules," in *International conference on machine learning* (CA, USA: PMLR), 2731–2741.

Hong, H., Zheng, L., Zhu, J., Pan, S., and Zhou, K. (2017). Automatic recognition of coal and gangue based on convolution neural network. Available at: https://arxiv.org/abs/1712.00720 (accessed on December 3, 2017).

Hou, Q. B., Zhou, D. Q., and Feng, J. S. (2021) "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 13708–13717. doi:10.48550/ar**v.2103.02907

Hu, F., Zhou, M., Yan, P., Liang, Z., and Li, M. (2022). A Bayesian optimal convolutional neural network approach for classification of coal and gangue with multispectral imaging. *Opt. Lasers Eng.* 156, 107081. doi:10.1016/j.optlaseng.2022.107081

Jiang, X., Pang, Y., Li, X., Pan, J., and Xie, Y. (2018). Deep neural networks with elastic rectified linear units for object recognition. *Neurocomputing* 275, 1132–1139. doi:10.1016/j.neucom.2017.09.056

Kim, D., Kim, J., and Kim, J. (2020). Elastic exponential linear units for convolutional neural networks. *Neurocomputing* 406, 253–266. doi:10.1016/j.neucom.2020.03.051

Koonce, B., and Koonce, B. (2021). ResNet 34. *Convolutional Neural Netw. Swift Tensorflow Image Recognit. Dataset Categ.*, 51–61. doi:10.1007/978-1-4842-6168-2_5

Kou, Q., Ma, H., Xu, J., Jiang, H., and Cheng, D. (2023). Coal flow foreign body classification based on ESCBAM and multi-channel feature fusion. *Sensors* 23 (15), 6831. doi:10.3390/s23156831

Leng, Z., Tan, M., Liu, C., Cubuk, E. D., Shi, X., Cheng, S., et al. (2022). A polynomial expansion perspective of classification loss functions. Available at: https://arxiv.org/abs/2204.12511 (accessed on May 10, 2022).

Li, Y., Wu, C. Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., et al. (2022). "Mvitv2: improved multiscale vision transformers for classification and detection," in *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, 4804–4814. doi:10.48550/arXiv.2112.0152

Liu, F., Liu, M., Zhang, L., and Wang, F. (2024). Foreign object classification method for coal conveyor belts based on residual networks (in Chinese). *Electron. Meas. Technol.*, 1–9. doi:10.19651/j.cnki.emt.2415997

Liu, Y., Wang, X., Zhang, Z., and Deng, F. (2023). Deep learning based data augmentation for large-scale mineral image recognition and classification. *Miner. Eng.* 204, 108411. doi:10.1016/j.mineng.2023.108411

Liu, Y., Zhang, Z., Liu, X., Wang, L., and Xia, X. (2021). Deep learning-based image classification for online multi-coal and multi-class sorting. *Comput. and Geosciences* 157, 104922. doi:10.1016/j.cageo.2021.104922

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., et al. (2022). "Swin transformer v2: scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019. doi:10.48550/arXiv.2111.09883

Ma, N., Zhang, X., Zheng, H. T., and Sun, J. (2018). "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S. W., Anwer, R. M., et al. (2022). "Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications," in *European conference on computer vision* (Cham: Springer Nature Switzerland), 3–20. doi:10.1007/978-3-031-25082-8_1

Müller, S. G., and Hutter, F. (2021). "Trivialaugment: tuning-free yet state-of-the-art data augmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 774–782.

Önal, M. K., Avci, E., Özyurt, F., and Orhan, A. (2020). "Classification of minerals using machine learning methods," in *2020 28th signal processing and communications applications conference (SIU) (IEEE)*, 1–4. doi:10.1109/SIU49456.2020.9302427

Pu, Y., Apel, D. B., Szmigiel, A., and Chen, J. (2019). Image recognition of coal and coal gangue using a convolutional neural network and transfer learning. *Energies* 12 (9), 1735. doi:10.3390/en12091735

Su, L., Cao, X., Ma, H., and Li, Y. (2018). "Research on coal gangue identification by using convolutional neural network," in *2018 2nd IEEE advanced information management, communicates, electronic and automation control conference (IMCEC) (IEEE)*, 810–814.

Tan, M., and Le, Q. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks," in *International conference on machine learning* (CA, USA: PMLR), 6105–6114.

Tan, M., and Le, Q. V. (2021). "Efficientnetv2: smaller models and faster training," in *International conference on machine learning*. PMLR 139, 10096–10106.

Touvron, H., Cord, M., and Jégou, H. (2022). "Deit iii: revenge of the vit," in *European conference on computer vision* (Cham: Springer Nature Switzerland), 516–533. doi:10.1007/978-3-031-20053-3_30

Zeng, C., Zheng, J., and Li, J. (2019). Real-time conveyor belt deviation detection algorithm based on multi-scale feature fusion network. *Algorithms* 12 (10), 205. doi:10.3390/a12100205

Zhang, J., Gao, Q., Luo, H., and Long, T. (2022b). Mineral identification based on deep learning using image luminance equalization. *Appl. Sci.* 12 (14), 7055. doi:10.3390/app12147055

Zhang, K., Wang, W., Lv, Z., Fan, Y., and Song, Y. (2021). Computer vision detection of foreign objects in coal processing using attention CNN. *Eng. Appl. Artif. Intell.* 102, 104242. doi:10.1016/j.engappai.2021.104242

Zhang, M., Cao, Y., Jiang, K., Li, M., Liu, L., Yu, Y., et al. (2022a). Proactive measures to prevent conveyor belt Failures: deep Learning-based faster foreign object detection. *Eng. Fail. Anal.* 141, 106653. doi:10.1016/j.engfailanal.2022.106653

Zhang, Q. L., and Yang, Y. B. (2021). "Sa-net: shuffle attention for deep convolutional neural networks," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP) (IEEE)*, 2235–2239. doi:10.1109/ICASSP39728.2021.9414568