Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.071880

ARTICLE

KPA-ViT: Key Part-Level Attention Vision Transformer for Foreign Body Classification on Coal Conveyor Belt

Haoxuanye Ji*, Zhiliang Chen, Pengfei Jiang, Ziyue Wang, Ting Yu and Wei Zhang

CCTEG Coal Mining Research Institute, Beijing, 100013, China

*Corresponding Author: Haoxuanye Ji. Email: jihaoxuanye@163.com

Received: 14 August 2025; Accepted: 01 October 2025; Published: 12 January 2026

ABSTRACT: Foreign body classification on coal conveyor belts is a critical component of intelligent coal mining systems. Previous approaches have primarily utilized convolutional neural networks (CNNs) to effectively integrate spatial and semantic information. However, the performance of CNN-based methods remains limited in classification accuracy, primarily due to insufficient exploration of local image characteristics. Unlike CNNs, Vision Transformer (ViT) captures discriminative features by modeling relationships between local image patches. However, such methods typically require a large number of training samples to perform effectively. In the context of foreign body classification on coal conveyor belts, the limited availability of training samples hinders the full exploitation of Vision Transformer's (ViT) capabilities. To address this issue, we propose an efficient approach, termed Key Part-level Attention Vision Transformer (KPA-ViT), which incorporates key local information into the transformer architecture to enrich the training information. It comprises three main components: a key-point detection module, a key local mining module, and an attention module. To extract key local regions, a key-point detection strategy is first employed to identify the positions of key points. Subsequently, the key local mining module extracts the relevant local features based on these detected points. Finally, an attention module composed of self-attention and cross-attention blocks is introduced to integrate global and key part-level information, thereby enhancing the model's ability to learn discriminative features. Compared to recent transformer-based frameworks—such as ViT, Swin-Transformer, and EfficientViT—the proposed KPA-ViT achieves performance improvements of 9.3%, 6.6%, and 2.8%, respectively, on the CUMT-BelT dataset, demonstrating its effectiveness.

KEYWORDS: Foreign body classification; global and part-level key information; coal conveyor belt; vision transformer (ViT); self and cross attention

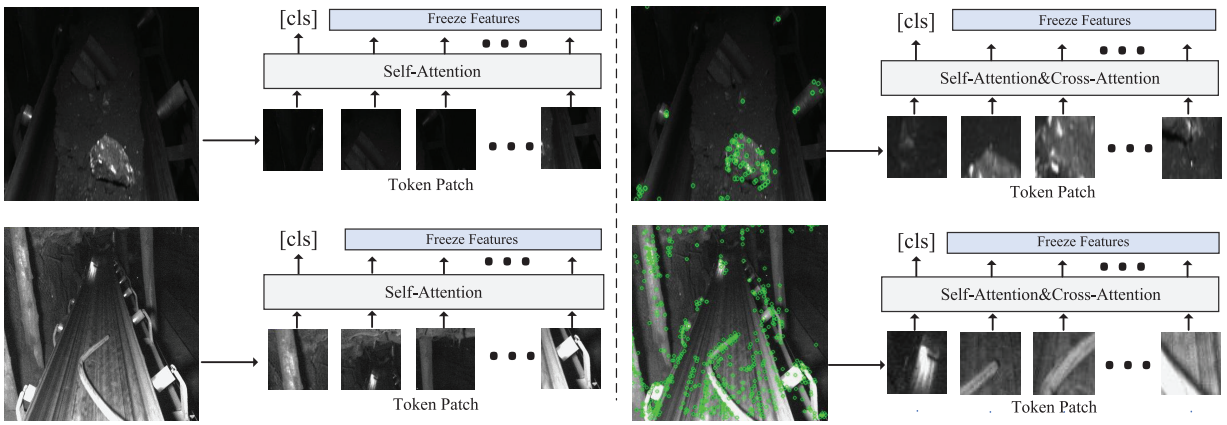
1 Introduction

Foreign body classification on coal conveyor belts is essential for the timely detection of foreign objects that may cause damage during high-speed operation. For instance, large gangue or anchor rods can scratch or tear the belt, while accumulated coal may lead to blockages at discharge points. It is therefore critical to accurately identify and categorize foreign bodies—such as oversized rocks and bolts—on the conveyor belt to enable early warnings and prevent potential hazards, thereby enhancing the safety of coal mining operations [1]. Nevertheless, the challenging roadway conditions, including poor illumination and complex background textures, often compromise classification accuracy. Traditional methods [2–4] are prone to misclassification, as they may incorrectly focus on irrelevant background features or struggle with low-light conditions [5].

Owing to the feature invariance and adaptable receptive fields of convolutional neural networks (CNNs), CNN-based frameworks have been widely adopted for foreign object classification tasks in this context [1,6,7]. For instance, reference [6] employed the VGG16 architecture [8] combined with transfer learning to develop a foreign body recognition model. Reference [7] enhanced the classic LeNet-5 network [9] to better accommodate the specific requirements of foreign object image identification. Meanwhile, reference [1] leveraged the translation invariance of convolution operations to integrate low-level spatial information into high-level semantic features via a residual network structure. Nevertheless, due to the insufficient modeling of correlations between local image patches, CNN-based approaches often constrain the expressive capacity of spatial information within features, leading to noticeable performance bottlenecks [10].

Inspired by the Non-Local method [11], preserving valid local information within key regions can significantly enhance the model's perceptual capabilities. Dosovitskiy et al. [12] introduced the Vision Transformer (ViT) framework, which learns correlations among pixels within patches to extract discriminative image features. However, ViT-based approaches such as [12,13] require large volumes of accurately labeled training images to learn effective representations; otherwise, they are prone to overfitting, limiting the model's generalizability and practicality in dynamic, complex real-world scenarios [10]. In the context of foreign body classification, the scarcity of extensive annotated data hinders ViT's ability to achieve strong discriminative performance. Thus, it is essential to extract meaningful training signals from the available dataset to facilitate and stabilize the learning process.

In the patch extraction process of ViT-based methods, a significant amount of irrelevant background information is often incorporated into the token vectors, which impedes the model's ability to quickly perceive semantically relevant features in the training images, as illustrated in Fig. 1a. To address this issue, an effective strategy involves detecting keypoints in each image using a dedicated keypoint detection algorithm and expanding these points into local patches within their surrounding regions. Compared to uniformly divided patches, these keypoint-centered patches contain richer and more discriminative information, as demonstrated in Fig. 1b.



(a) Traditional patch partitioning strategy (b) Keypoint-based patch partitioning strategy

Figure 1: The training samples and their corresponding patches are segmented using (**a**) a conventional patch partitioning strategy and (**b**) a key-point-guided patch partitioning strategy

To fully leverage the semantic information within these patches, we employ self-attention and cross-attention modules to integrate key local features into the image representation for foreign object classification. The effectiveness of this key information is validated on the coal conveyor belt dataset CUMT-BelT [1]. Furthermore, to assess the generalization capability of our approach, we conduct additional

evaluations on common classification benchmarks, namely the CIFAR-10 and CIFAR-100 datasets [14]. Experimental results demonstrate that our method achieves competitive performance compared to other state-of-the-art methods.

The contributions of our paper are summarized as follows:

- This paper proposes a novel Key Part-level Attention Vision Transformer (KPA-ViT) framework for foreign body classification on coal conveyor belts, aiming to mitigate overfitting caused by insufficient effective training information.
- This paper proposes a key local mining strategy that detects discriminative key points based on image appearance characteristics and extracts informative local regions to facilitate model training.
- This paper introduces a novel attention module designed to effectively integrate the extracted key local features with global features, thereby enabling more comprehensive utilization of both local and global information.

The remainder of this paper is structured as follows: Section 2 offers a concise summary of pertinent research. Section 3 introduces the proposed transformer framework, which leverages global and local key information for foreign object classification on coal conveyor belts. Section 4 presents experimental results and discussions aimed at validating the proposed method's efficacy. Lastly, Section 5 presents the conclusion.

2 Related Work

The section briefly reviews related work on foreign body classification methods, as well as the feature extraction frameworks.

2.1 Foreign Body Classification

The objective of foreign body classification is to quickly identify images containing foreign bodies in industrial processes, which is an important part of industrial warning. The traditional methods use image prior-knowledge feature fusion and general classification decision technology to achieve this task. Wang et al. [2] use the inter-frame difference method, threshold classification, and select-shape operators to identify large foreign bodies in belt conveyors. He et al. [3] use support vector machine technology to classify foreign bodies in combination with their texture and gray features. Reference [4] combines the K-nearest neighbor algorithm, support vector machine, and multi-feature fusion technology for foreign body recognition. However, since the prior knowledge is greatly dependent on the clear appearance of the image, these methods as a whole have problems such as poor robustness and susceptibility to illumination.

Relying on the development of deep learning technology, the deep learning framework is used to adapt to the classification of foreign bodies in different scenarios through continuous iterative calculation. The method based on a deep convolutional network extracts class-specific features through a large-scale parameter network and then determines whether the image contains foreign objects through the decision level [6]. Whata et al. [15] insert the uncertainty quantification techniques into the Bayesian Neural Networks and the Deep Neural Network to solve the multi-class classification of chest X-ray images. Amaya et al. [16] propose an iterative training process that aims to identify a representative data subset and a novel distance-based kernel based on a similarity matrix, leading to improved inferences about the population to solve both binary and multi-class classification problems. These CNN-based methods are easy to optimize and can achieve reliable classification results. However, these methods have insufficient perception of the local characteristics of the targets in each sample. They are easy to be misled by complex background information in images, thus limiting the classification performance.

2.2 Feature Extraction Framework

The feature extraction framework plays an important role in computer vision tasks. Most previous methods typically extract global feature representation or object-specific voxel information per image for image classification [17,18]. When the background is cluttered or the pedestrian is occluded, part-level features can improve the performance. Non-local neural network [11] utilizes the relationships between pixel-aware features to explore the effective object information, thus further enhancing the instance-level discriminative capability of extracted features per image.

Inspired by local information interaction [11], ViT [12] directly encodes local patches in an image, and the information between these encodings is interacted with by a self-attention mechanism. Based on this, DETR [19] and BEVFormer [20] introduced the cross-attention mechanism [21] to learn the target feature in each query by grouping the local feature information from the image database into a set of questions and mapping the matching target. Serikbay et al. [22] successfully apply deep convolutional neural networks (CNNs) to classify high-voltage insulator surface conditions. H-QNN [23] combines quantum and classical neural networks, which use a two-qubit quantum circuit with a classical convolutional architecture to enhance classification accuracy and address the overfitting for small datasets. PBNet [24] combines a CNN and a Transformer complementary network to address the issues of weakly textured images. DCD-GCN [25] proposes a dual-channel deep graph convolutional neural network, which applies residual connections to dual-channel graph convolutional neural networks, and increases the depth of dual-channel graph convolutional neural networks.

3 Methods

Given a labeled dataset $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^{3\times H\times W}$ is the i-th training image and N is the number of training images, H and W denote the height and width of each training image x_i; y_i is the class label of x_i. The goal of foreign body classification on coal conveyor belts is to train a model based on $\mathcal{X}$ that can accurately classify the types of foreign bodies in each input image. The KPA-ViT framework is presented for the task in Fig. 2. The framework consists of three components: a key local mining method is to search the part-level patch by the detected key points of each x_i; an attention module uses self-attention and cross-attention modules to integrate and enhance global and part-level features, and predicts the final classification results through a decision module.

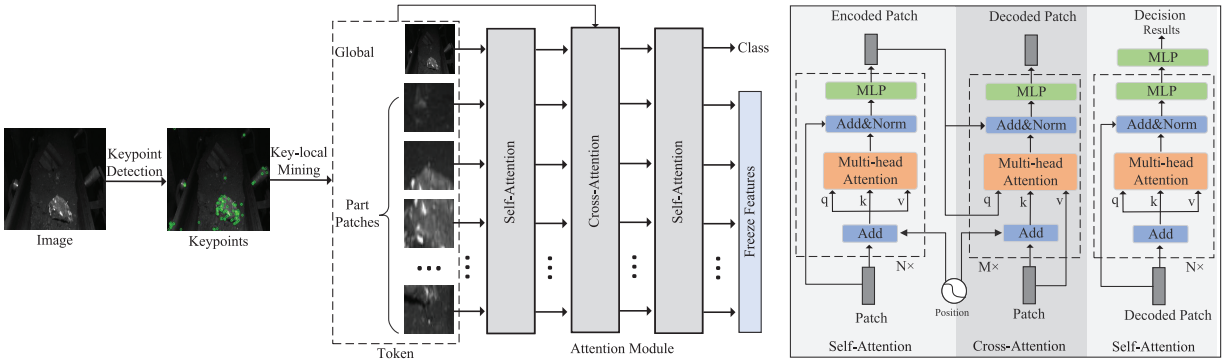


Figure 2: Overview of our framework. For a query image, we first utilize a key-point detector to generate the key-point positions. Then, the framework is to mine key locals and filter redundant locals. Finally, these mine key locals are introduced into the proposed attention module to calculate the classification results by self-attention and cross-attention blocks

3.1 Key Local Mining

For each input image x_i, the key point detection method is to calculate the location of its key points using the appearance characteristics of x_i. For convenience, we adopt the traditional ORB (Oriented Fast and Rotated Brief) key point detection method [26], which first uses the FAST (Features from Accelerated Segment Test) method [14] to quickly detect corner points, and then uses the Harris corner measurement [27] to screen N points with the largest Harris response value from the extracted FAST corner points as the key points of x_i. These key points of x_i are denoted as $\mathcal{P}_i = \{p_k\}_{k=1}^{K}$, where p_k denotes the k-th key point of x_i.

After obtaining the key points $\mathcal{P}_i$, the key local mining strategy can mine the key part-level patches for the attention module. However, some key points come together. Therefore, we should filter the key region corresponding to each key point to reduce the key local redundancy. The key locals' filter process is formulated as:

$$\begin{aligned} \mathcal{L}_k &= \{\mathrm{L}(p_j) | \mathrm{IOU}(\mathrm{L}(p_j), \mathrm{L}(p_k)) < \mu\}_{j=1}^{K} \\ \mathcal{L}_i &= \{\mathcal{L}_k\}_{k=1}^{K}, \end{aligned} \tag{1}$$

where, $\mathrm{L}(p_j)$ represents a $N' \times N'$ image neighborhood with p_j as the center point. $\mathrm{IOU}(a, b) = (a \cap b)/(a \cup b)$ means the IOU (intersection-over-union) ratio of set a and b. μ is an IOU hyper-parameter, which is used to select the key locals. $\mathcal{L}_k$ represents the key locals based on the k-th key point, and $\mathcal{L}$ denotes the key locals of x_i. The key local generation procedure is presented in Algorithm 1. In this way, the patches based on the mining of key textures in the image contain more category-relevant appearance/texture information compared to the original patches without background filtering.

Algorithm 1: Key local generation procedure

Require: Training image x_i, IOU threshold μ, Patch size N'.
Ensure: Generated key locals $\mathcal{L}_i = \{\}$,
1: **Init.**: Key points $\mathcal{P}_i = \{\}$, Key locals $\mathcal{L}_i = \{\}$;
2: Obtaining Key points $\mathcal{P}_i$ of x_i by FAST method, i.e., $\mathcal{P}_i \leftarrow \mathrm{FAST}(x_i)$;
3: **for** p_j in $\mathcal{P}_i$ **do**
4: Cropping patches by p_j and N', i.e., $L(p_j) = x_i[p_j - (\frac{N'}{2}, \frac{N}{2}) : p_j + (\frac{N'}{2}, \frac{N'}{2})]$;
5: **end for**
6: **for** p_j in $\mathcal{P}_i$ **do**
7: **for** p_k $(k \neq j)$ in $\mathcal{P}_i$ **do**
8: $\mathrm{IOU}_j \leftarrow \mathrm{IOU}(L(p_j), L(p_k))$
9: **end for**
10: $\mathcal{L}_i \leftarrow L(p_j)$ If $\mathrm{IOU}(L(p_j), L(p_k)) < \mu$
11: **end for**
12: **return** Generated key locals $\mathcal{L}_i$

3.2 Attention Module

After key local mining, we introduce these key local $\mathcal{L}$ with rich key local information of x_i into the attention module, thus expanding the training information in the samples to avoid overfitting in training and further boosting the discriminative information in the features. The attention module consists of four components, namely the embedding block, self-attention block, cross-attention block, and decision block. We will cover the technical details of each component in detail below.

Embedding block. The embedding block is to generate the global and part-level patch embedding of each x_i. We use convolutional neural networks to encode these global and local patches to retain the global and key local texture cues of x_i, rather than MLP (Multilayer Perceptron) layers in the ViT framework [12]. First, the embedding block uses the convolution layer of $1 \times 1 \times C$ to increase the number of channels in the image, and uses a convolution of $N' \times N' \times C$ to generate a patch embedding for each patch. The process is formulated as:

$$\mathbf{E} = \left[\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_{|\mathcal{L}|}\right], \tag{2}$$

where $\mathbf{e}_0 \in \mathbb{R}^C$ is a global patch embedding, which first resizes x_i as $N' \times N'$ by a convolutional layer and is generated by an embedding block. $\mathbf{e}_1 \in \mathbb{R}^C$ denotes the key part-level embedding of each $\mathrm{L}(p_j)$ in $\mathcal{L}$.

Self-attention block. The self-attention module encodes each patch embedding in $\mathbf{E}$, which is used to capture the long-term dependency relationships between the global and part-level embedding tokens in $\mathbf{E}$. First, we assign an initial learnable position embedding for each patch embedding in $\mathbf{E}$, which is denoted as $\mathbf{E}' = [\mathbf{e}'_0, \mathbf{e}'_1, \ldots, \mathbf{e}'_{|\mathcal{L}|}]$. Then, we introduce the position embedding $\mathbf{E}'$ in each embedding in $\mathbf{E}$ by adding operation, i.e., $\mathbf{E} + \mathbf{E}' = [\mathbf{e}_0 + \mathbf{e}'_0, \mathbf{e}_1 + \mathbf{e}'_1, \ldots, \mathbf{e}_{|\mathcal{L}|} + \mathbf{e}'_{|\mathcal{L}|}]$. Following ViT [12], we insert each $\mathbf{e}_i + \mathbf{e}'_i$ in $\mathbf{E}$ into q, k and v, i.e., $q = k = v \leftarrow \mathbf{e}_i + \mathbf{e}'_i$, where q, k, v are the query, key and value vectors, respectively. And then, q, k, v are fed into the multi-head attention to encode each part local $\mathrm{L}(p_i)$ in $\mathcal{L}$, as:

$$\begin{aligned} &\mathrm{MultiHead}(q,k,v) = \mathrm{Concat}(h_1, \ldots, h_n) \cdot \mathbf{W}^o, \\ &h_i = \mathrm{Softmax}\left[\frac{(q \cdot \mathbf{W}_i^q) \cdot (k \cdot \mathbf{W}_i^k)^{\mathrm{T}}}{\sqrt{d_k}}\right] \cdot (v \cdot \mathbf{W}_i^v), \end{aligned} \tag{3}$$

where MultiHead() denotes the multi-head attention operation, Concat() denotes the vector concatenating attention operation. h_i is the output vector extracted by the i-th attention head. $\mathbf{W}_i^q$, $\mathbf{W}_i^k$ and $\mathbf{W}_i^v$ denote three non-shared learnable weights. Softmax is an activation function, which is used to normalize the feature distribution. $d_k = C/n$ denotes the dimension of the output vector by the i-th attention head. $\mathbf{W}^o$ is used to integrate the vector h_i output by each attention head.

Finally, in order not to reduce the feature information in each patch, we employ the residual strategy to mix each $\mathbf{e}_i + \mathbf{e}'_i$ and $\mathrm{MultiHead}(q,k,v)$, which is formulated as:

$$\mathbf{e}_i^* = \mathrm{MLP}(\mathrm{L_2Norm}(\mathbf{e}_i + \mathbf{e}'_i + \mathrm{MultiHead}(q,k,v))), \tag{4}$$

where $\mathbf{e}_i^* \in \mathbb{R}^C$ is the encoded embedding by the self-attention module. $\mathrm{L_2Norm}()$ means a L_2 normalization operation, MLP() is Multi-layer Perceptron (MLP).

Cross-attention block. To fully aggregate the global and key part-level information, inspired by DETR [19], we design a simple yet effective cross-attention block to decode each $\mathbf{e}_i^*$. Thus, the multi-head attention of cross-attention is formulated as:

$$\begin{aligned} &\mathrm{MultiHead}(\mathbf{e}_i^*,k,v) = \mathrm{Concat}(h_1, \ldots, h_n) \cdot \mathbf{W}^{do}, \\ &h_i = \mathrm{Softmax}\left[\frac{(\mathbf{e}_i^* \cdot \mathbf{W}_i^{dq}) \cdot (k \cdot \mathbf{W}_i^{dk})^{\mathrm{T}}}{\sqrt{d_k}}\right] \cdot (v \cdot \mathbf{W}_i^{dv}), \end{aligned} \tag{5}$$

where $\mathbf{W}_i^{dq}$, $\mathbf{W}_i^{dk}$ and $\mathbf{W}_i^{dv}$ denote the learnable parameters of $\mathbf{e}_i^*$, k and v in the cross-attention block. $\mathbf{e}_i^*$ includes the more aggregation cues of global and key part-level. By the cross-attention block, the information of $\mathbf{e}_i^*$ can be introduced in the generated decoded embedding $\mathbf{e}_i^d$, as:

$$\mathbf{e}_i^d = \mathrm{MLP}(\mathrm{L_2Norm}(\mathbf{e}_i + \mathbf{e}'_i + \mathrm{MultiHead}(\mathbf{e}_i^*,k,v))). \tag{6}$$

As a result, the cross-attention block is to fuse the self-attention and original embedding information, and through attention weights to achieve selective transmission of information, it is used to generate better image descriptors.

Decision block. The goal of the decision block is a classifier to predict the category of each x_i. To better understand the key information in $\mathbf{e}_i^d$, as shown in Fig. 2, the self-attention blocks are exploited. And then, an MLP is used to predict the category y_i^* of each x_i. The decision block is formulated as:

$$\mathbf{z}_i = \mathrm{MLP}(\mathrm{L_2Norm}(\mathbf{e}_i^d + \mathrm{MultiHead}(\mathbf{e}_i^d, \mathbf{e}_i^d, \mathbf{e}_i^d))). \tag{7}$$

Finally, the category probability distribution is generated by $\mathbf{z}_i = \mathrm{MLP}(\mathbf{z}_i)$, where $\mathbf{z}_i \in \mathbb{R}^{|\mathrm{class}|}$. $|\mathrm{class}|$ is the category number. $y_i^* = \mathrm{argmax}\ \mathbf{z}_i$ is the predicted category of x_i. The attention procedure is shown in Algorithm 2.

Algorithm 2: Attention procedure

Require: Training image x_i, Key locals $\mathcal{L}_i$, The number of Self-Attention blocks N, The number of Cross-Attention blocks M, The number of Decision blocks N.
Ensure: Category probability distribution $\mathbf{z}_i$
1: **Init.**: Position embedding $\mathbf{P}$;
2: Obtaining $\mathbf{e}_0$ of x_i and $\{\mathbf{e}_j\}_{j=1}^{|\mathcal{L}|}$ in $\mathcal{L}_i$ by Embedding block;
3: **for** $t = 1, t \leq N, t++$ **do** # Self-attention block
4: Obtaining the encoded embedding $\mathbf{e}_i^*$ by Eq. (4);
5: **end for**
6: **for** $t = 1, t \leq M, t++$ **do** # Cross-attention block
7: Generating the encoded embedding $\mathbf{e}_i^d$ with $\mathbf{e}_i$ and $\mathbf{e}_i^*$ by Eq. (6);
8: **end for**
9: **for** $t = 1, t \leq N, t++$ **do** # Decision block
10: Generating the decision embedding $\mathbf{z}_i$ by Eq. (7);
11: **end for**
12: Obtaining the category probability distribution $\mathbf{z}_i \in \mathbb{R}^{|\mathrm{class}|}$ by $\mathbf{z}_i = \mathrm{MLP}(\mathbf{z}_i)$;
13: **return** $\mathbf{z}_i$

3.3 Loss Function

We optimize the foreign body classification framework by maximizing $\mathbf{z}_i[y_i]$. We exploit the cross-entropy loss function as:

$$\mathrm{Loss} = -\frac{1}{|\mathcal{B}|}\sum_{x_i \in \mathcal{B}} \log \frac{\exp(\mathbf{z}_i[y_i])}{\sum_{y_j} \exp(\mathbf{z}_i[y_j])}, \tag{8}$$

where $\mathcal{B}$ is a mini-batch samples and $|\mathcal{B}|$ is the mini-batch size. The overall optimization procedure is presented in Algorithm 3.

Algorithm 3: Optimization procedure

Require Training dataset $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^N$, initialized transformer-based framework Q, training epoch T.
Ensure Optimized Q,
1: **Init.**: Detection key points $\mathcal{P}$;
2: **for** $t = 1, t \leq T, t++$ **do**

(Continued)

Algorithm 3 (continued)

```
    for each x_i in 𝒳 do
        Generation ℒ for x_i by 𝒫_i;
        # Patch embedding encoding
        Generation patch embedding E by Eq. (4);
        Encoding each patch embedding in E by self-attention block;
        # Patch embedding decoding
        Decoding each patch embedding by a cross-attention block with the encoded patch embedding;
        # Category decision
        Calculating the discriminative features of each e_i^d by a self-attention;
        Prediction of the category by an MLP;
        # Discriminative feature learning
        Optimize Q by minimizing the loss function Eq. (8);
    end for
end for
return Optimized Q
```

4 Experiments and Discussions

4.1 Datasets and Protocol

Datasets. We evaluate our method on a foreign body classification on the coal conveyor dataset, i.e., CUMT-BelT [1]. Moreover, to verify the effectiveness of our proposed method on the public classification task, we also evaluate the method on cifar-10 [14] and cifar-100 [14] datasets.

- **CUMT-BelT** dataset [1] collects from the underground belt transport environment, a total of 6000 images, divided into three categories, namely block, bolt, and normal pictures. Each category has 2000 images, including 1600 training images and 400 test images.
- **Cifar-10** dataset [14] consists of 60,000 color images in 10 categories, 6000 images in each category, 5000 images in the training dataset, and the rest in the test.
- **Cifar-100** dataset [14] consists of 60,000 color images containing 100 categories, 600 images per class, 500 images in the training dataset, and 100 images in the test dataset.

Protocol. Following the common experimental setting in [1,14], we employ the classification accuracy to evaluate the performance of the methods. The classification accuracy is the proportion of samples correctly predicted by the category to all samples, i.e., acc. = $\sum_{x_i \in \mathcal{X}_{\text{test}}} \mathrm{I}(y_i^* == y_i)/|\mathcal{X}_{\text{test}}|$. $\mathrm{I}(a)$ is a logical function, $\mathrm{I}(a) = 1$ only a is True. $\mathcal{X}_{\text{test}}$ is the testing dataset, and $|\mathcal{X}_{\text{test}}|$ means the cardinality of $\mathcal{X}_{\text{test}}$.

4.2 Implementation Details

Our method is implemented in PyTorch [28] with two NVIDIA GeForce RTX 4090 Ti. We process the training images with a random crop and a random flip. During optimization, our framework is optimized by the SGD [29] method with a learning rate of 0.001 and the mini-batch size $|\mathcal{B}| = 8$ following the ViT fine-tuning process. We train the method with 100 epochs in total and adopt a cosine learning rate scheduler strategy following ViT [12]. About the key local mining, to balance the redundancy and richness of the training information, we limit the $\mu \in \{0.1, 0.2, \ldots, 0.9\}$ and select the μ corresponding to the best classification accuracy, which is discussed in Section 4.3 "IOU threshold μ", and we set the IOU threshold μ as 0.7. The key local size N' is setting as 16 to generate 16×16 patches for our KPA-ViT framework, and we also discuss N' in Section 4.3 "Key local N'". To avoid the time-consuming grid search process, we set $M = N$.

According to the number of attention heads in ViT-base being 12, the total number of attention heads in the self-attention and cross-attention blocks is set as $N + M + N = 12$. Therefore, we set $M = N = 4$. Following VIT [12], the number position embedding is set as 196, and we randomly select 196 key locals in $\mathcal{L}$.

4.3 Ablation Study

To evaluate the effectiveness of each component in our proposed method, experiments are conducted as follows.

1. "Base" denotes only self-attention blocks for the classification task;
2. "Base+random" denotes the "Base." with randomly selection locals;
3. "Base+key" denotes introducing the key locals into "Base.";
4. "Base+key+cross" is inserting the cross-attention blocks into "Base+key";
5. "Base+key+cross+decision" incorporates the decision blocks into "Base+key+cross";
6. "Base+cross+decision" removes the key locals from "Base+key+cross+decision".

To align the experimental settings, the attention head numbers of each component are the same, i.e., 12. Table 1 shows the experimental results.

Table 1: Ablation study about the effectiveness of the key local mining, the cross-attention, and decision blocks in our method on CUMT-BelT, Cifar-10, and Cifar-100. The best is in **bold**

Components	Class Acc. (%)			
	CUMT-BelT	Cifar-10	Cifar-100	Mean
Base	82.0	90.9	66.5	79.8
Base+random	81.9	90.2	67.1	79.7
Base+key	84.2	91.6	69.4	81.7
Base+key+cross	91.1	94.5	76.5	87.4
Base+key+cross+decision	**91.8**	**95.3**	**76.7**	**87.9**
Base+cross+decision	88.1	92.4	70.3	83.6

The effectiveness of key local mining. As shown in Table 1 "Base+random" has similar performances as "Base" on CUMT-BelT, Cifar-10, and Cifar-100. The results indicate that the randomly selected locals do not have effective training information. After incorporating the mined key locals, the results are presented in Table 1 "Base+key". We can observe that the classification accuracy of the "Base+key" is improved by 2.2%, 0.7%, and 2.9% over the "Base" on CUMT-BelT, Cifar-10, and Cifar-100, respectively. Also, the mean classification accuracy on the three datasets of "Base+key" is improved by 1.9% and 2.0% than "Base" and "Base+random". This is because the mined key locals can provide more effective information than the uniform and random selection patches, which can increase the distinctiveness of the features to boost the classification results.

The effectiveness of cross-attention blocks. Table 1 demonstrates that the classification results of "Base+key+cross" are improved by 6.9%, 2.9%, and 7.1% over "Base+key" on CUMT-BelT, Cifar-10, and Cifar-100, which determines the effectiveness of the cross-attention blocks. After incorporating the cross-attention blocks, the mean performance is improved by 5.7% and 7.6% over "Base+key" and "Base". The reason is that the cross-attention blocks can fully aggregate the key part-level information into the patch embedding to help the model distinguish the differences of categories better.

The impact of decision blocks. The experimental results of the decision blocks are presented in Table 1 "Base+key+cross+decision". Compared to the component "Base+key+cross", the classification results are higher at 0.7%, 0.8%, and 0.2%, which indicate the effectiveness of decision blocks after the cross-attention block. Introducing the decision blocks can improve the mean classification accuracy 0.5% on the three datasets than "Base+key+cross". Moreover, we also evaluate the component "Base+cross+decision" which removes the key local from the overall framework. We can observe that the performances have both significantly dropped. The results present that the key part-level information plays an important role in foreign body classification tasks.

4.4 Parameter Analysis

We explore the impacts of the IOU threshold μ, the key local size N', and the number of attention heads N in self-attention blocks and M in cross-attention blocks on the CUMT-BelT.

IOU threshold μ. According to the description of Eq. (1), μ is used to filter out the locals with a high repetition rate to avoid redundant information, thus ensuring the diversity of information in samples. Fig. 3a reports the classification accuracy by varying μ from 0.1 to 0.9 every 0.2. It can be seen that the performance is best at $\mu = 0.7$. A low μ creates redundant key locals, which dilute the diversity of key local information; while a higher μ reduces the number of key locals, thus affecting the effective training information in samples.

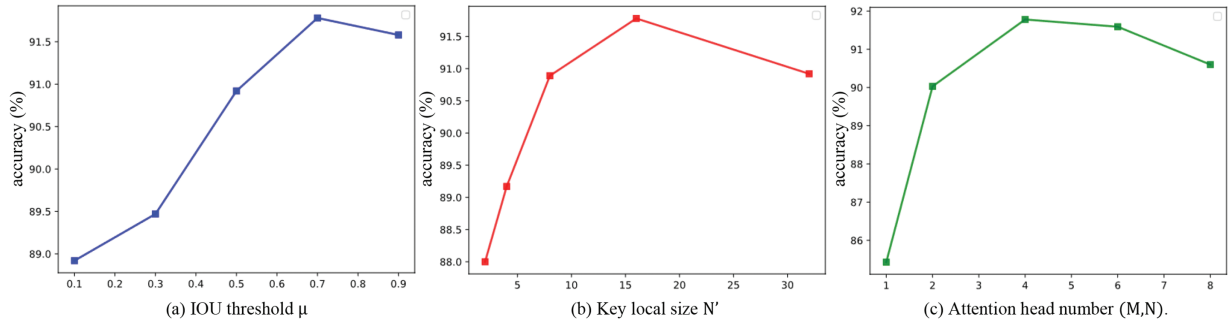


Figure 3: Experiments on the impact of (**a**) the IOU threshold μ, (**b**) the key local size N', and (**c**) the attention head number (M, N) on the CUMT-BelT dataset

Key local size N'. In the key local mining, the hyper-parameter N' affects the part-level information of each patch. The results are presented in Fig. 3b. It demonstrates that both classification accuracies increase significantly over the $N' = 16$ and then slightly decrease. A smaller N' will make the information in each patch not enough, while a larger N' will make the information in each patch hard to learn. Therefore, choosing a suitable N' is important for the overall performance, and we set $N' = 16$.

Attention head number M and N. M and N are the number of attention heads in the cross-attention blocks and the self-attention blocks, respectively. Since M and N are both used in the attention module to enhance the perceptibility of our method in the training and inference process, we set $M = N$ to reduce the parameter search difficulty. Thus, we can denote M and N as (M, N) for simplicity. We draw the classification accuracy curves by varying (M, N) in $\{1, 2, 4, 6, 8\}$, as given in Fig. 3c. The curves show that our method works well under different weights, which demonstrates the flexibility of our method. Specifically, when $(M, N) \in \{1, 2, 4\}$, the classification accuracy is increased on the CUMT-BelT dataset; when $(M, N) \in \{4, 6, 8\}$, both classification accuracy has a slight degradation in general. We speculate that smaller (M, N) will make the model too simple, thus the discriminative information in the samples is not perceived well,

while larger (M, N) will make of model contain more parameters, which are hard to optimize. Thus, we empirically choose $M = N = 4$ for our method.

4.5 Comparison Results

We proceed to compare our proposed method with the previous foreign body classification methods on coal conveyor belts, i.e., CUMT-BelT. Moreover, to verify the generalization of our method, we evaluate our method on two public classification benchmarks, i.e., Cifar-10 and Cifar-100. Table 2 presents the experimental results on CUMT-BelT [1], Cifar-10 [14], and Cifar-100 [14].

Table 2: Performance comparison with state-of-the-art foreign body classification methods on CUMT-BelT, Cifar-10 and Cifar-100 datasets, and the parameters and FLOPs (Floating Point Operations) on CUMT-BelT

	Methods	Class Acc. (%)			Parameters	FLOPs
		CUMT-BelT	Cifar-10	Cifar-100		
CNN-based	Shufflenet [30]	80.1	88.4	68.3	2.40 M	0.14 G
	MobileNetV2 [31]	80.9	89.1	69.1	2.30 M	2.01 G
	LeNet-5_Su [7]	77.7	84.9	66.3	6.30 M	3.44 G
	GoogleNet [32]	81.3	88.2	69.7	6.80 M	1.80 G
	VGG16 [8]	80.4	88.9	69.1	134.36 M	15.48 G
	ResNet34 [33]	82.9	92.1	71.1	22.30 M	3.78 G
	ResNet50 [33]	84.4	93.5	73.4	27.20 M	4.12 G
	ResNeXt50 (32*4d) [33]	84.6	94.0	73.8	26.8 M	4.04 G
	W_ResNet50 [2]	84.8	93.9	73.6	26.7 M	4.39 G
	ResNet110 [33]	84.3	94.1	73.8	53.80 M	8.20 G
	Cheng et al. [1]	85.1	94.1	73.9	16.70 M	2.98 G
Transformer-based	ViT [12]	82.0	90.9	66.5	85.65 M	13.49 G
	Swin-Transformer [13]	86.8	94.3	76.1	88.01 M	53.11 G
	EfficientViT [34]	91.3	**96.7**	78.5	12.43 M	9.21 G
	KPA-ViT	91.8	95.3	76.7	85.65 M	13.50 G
	KPA-ViT(S)	93.4	95.7	77.6	88.01 M	53.12 G
	KPA-ViT(E)	**94.1**	**96.7**	**78.8**	12.43 M	9.21 G

Note: The best is in **bold**.

On CUMT-BelT. The CUMT-BelT dataset is collected from real coal mine scenarios. Thus, our method mainly focuses on the benchmark. We first compare our method to the state-of-the-art foreign body classification methods. The results of our method on CUMT-BelT are presented in Table 2 "CNN-based", denoted as "CUMT-BelT". Compared to the CNN-based foreign body classification methods, the performance of our KPA-ViT surpasses that of these methods on the CUMT-BelT dataset [1]. The results determine that the effectiveness of our KPA-ViT is owing to the prior extraction of local texture information and the focus on semantic information in key regions compared with the CNN-based methods.

About the transformer-based framework that performs well on computer vision tasks, our method is more suitable for the foreign body classification task on the coal conveyor belt, with a result of is higher by 9.8% than the ViT framework. Moreover, compared to the recent transformer-based frameworks, such as Swin-transformer [13] and EfficientViT [34], our KPA-ViT method also achieves better results on the foreign body classification on coal conveyor belt, i.e., CUMT-BelT. To explore the effectiveness of the key-local information in the classification task, we also introduce this information into the swin-transformer and efficientViT frameworks, termed as KPA-ViT(S) and KPA-ViT(E). To suit the methods, we first reconstruct the token embedding in the KPA-ViT according to the location of each keypoint. And then, the multi-head attention blocks in self-attention and cross-attention modules are replaced as the attention blocks

in the Swin-transformer [13] and EfficientViT [34].The results are shown in Table 2 "Transformer-based". Due to the limitation of training data, these transformer-based foreign body classification methods cannot effectively address the interference of complex background information, which affects the foreign body classification results. However, our KPA-ViT method can quickly capture effective regions and ignore the impact of background information on features, thereby reducing the dependence on large-scale training data.

4.6 More Discussions

With deformable attention. Without losing generality, we explore our method with different attention heads to extract features for foreign body classification. We replace each attention head in the attention module with a deformation attention head in our method, which is called "Ours (Defor.)", and then evaluate it on CUMT-BelT, Cifar-10, and Cifar-100. The results are presented in Table 3. We can observe that the performances are slightly improved after incorporating the deformable attention module. The results determine that deformable attention is more suitable for foreign body classification. However, due to the more FLOPs (floating point operations) of deformation attention, i.e., 15.8 G, we do not adopt deformable attention, and it will be our further work.

Table 3: Performances of our method with deformable attention on CUMT-BelT, Cifar-10, and Cifar-100

Components	Class Acc. (%)		
	CUMT-BelT	**Cifar-10**	**Cifar-100**
ViT	82.0	90.9	66.5
Ours	91.8	**95.3**	76.7
Ours (Defor)	92.6	**95.3**	**77.1**
Ours (CNN+Trans)	**93.4**	94.9	76.9

Note: The best is in **bold**.

With CNN+Transformer. The CNN-transformer framework has achieved excellent performance in the object detection task, which generates a descriptor for each receptive field through the CNN framework, and then generates the corresponding image descriptor through the attention mechanism in the transformer. In order to combine the proposed method with the CNN-transformer framework, we retain the center points corresponding to the key regions. By sending the image into the CNN to generate the corresponding feature map, and then searching for the corresponding embedding tokens based on the center point coordinates. The corresponding category distribution is generated by the proposed attention module. The results are shown in Table 3 "CNN+Trans". We can observe that the performances of our method with CNN+Transformer, i.e., Ours (CNN+Trans), are slightly improved by 1.6% than "Ours" on the CUMT-BelT dataset. The CNN can provide more spatial context information for the embedding tokens, while the key region mining module can offer richer category-related clues for the image descriptors.

Training/inference speed analysis. We also evaluate the training and testing speed of our method on the CUMT-BelT dataset with two NVIDIA GeForce RTX 4090Ti. For our method, the training speed is 45.8 FPS and the testing speed is 120.4 FPS. In practical applications, the focus is on test speed. Therefore, our method is suitable for real-time foreign body classification tasks on coal conveyor belts.

Visualization results. To intuitively evaluate the effectiveness of our method, we visualize the features of the samples on the testing sets of CUMT-BelT learned by ViT, as in Fig. 4a, and our method, as in Fig. 4b. The points of the same color indicate that these samples are of the same class. The visualization results also indicate that our method can more accurately bring the training samples with the same class together than

the ViT framework, which shows the discriminative capability of the features helped by our method, thus validating the effectiveness of our key part-level information and the cross-attention module.

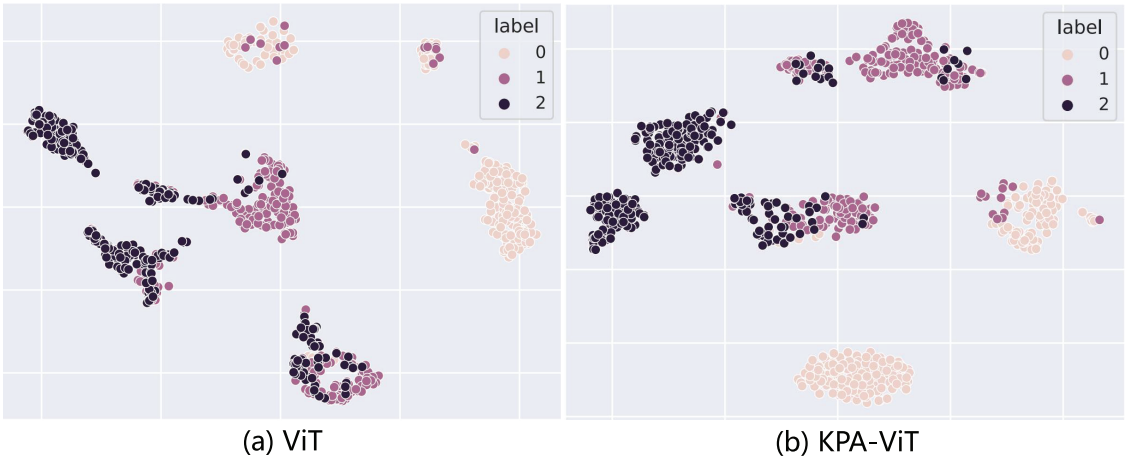


Figure 4: The t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization of the features of the testing samples on CUMT-BelT, which are extracted by (**a**) ViT and (**b**) KPA-ViT frameworks

5 Conclusion

In this paper, we propose a method for foreign body classification on a coal conveyor belt. Unlike most previous methods that mainly focus on global information and ignore key part-level information, the proposed method uses key local features with abundant appearance cues to boost the classification results. Specifically, the method first uses key-point detection to detect key points in images and then extracts the key part-level information by the key local mining strategy. Then, by introducing the attention module with cross-attention blocks to fully use the key part-level information for aggregating the features for foreign body classification. Extensive experimental results demonstrate that the proposed method is more suitable for the foreign classification on the coal conveyor belt, i.e., CUMT-BelT. Moreover, the method also has good results in the classification task, i.e., Cifar-10, and Cifar-100 datasets.

Limitations and Future Work. To fully explore the performance of our KPA-ViT, we present some incorrect and correct classification results with the keypoints, in Fig. 5. These visualization results determine that the keypoint detection may provide the erroneous keypoint locations caused by the lack of lighting conditions and complex background, etc. As a result, the local features corresponding to these erroneous keypoints always contain the key information of non-target objects (i.e., background), misleading the Attention Module to provide the incorrect classification results.

To sum up, a robust keypoint detection method can provide accurate keypoint locations in the absence of lighting and complex underground scenes, thus providing reliable key-local information for the Attention Module to boost the classification results.

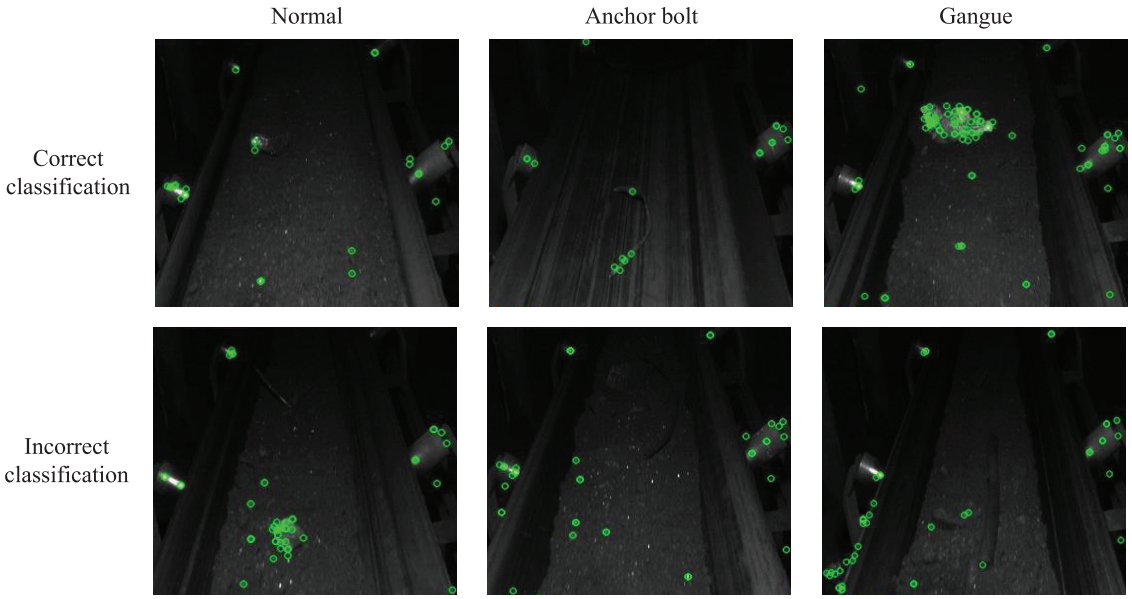


Figure 5: Some visualization results of the correct (upper) and incorrect (under) classification. From left to right, each column represents the image is normal (class:0), with an anchor bolt (class:1), and with Gangue (class:2)

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Key Research and Development Program of China (grant number 2023YFC2907600), the National Natural Science Foundation of China (grant number 52504132), Tiandi Science and Technology Co., Ltd. Science and Technology Innovation Venture Capital Special Project (grant number 2023-TD-ZD011-004).

Author Contributions: Conceptualization: Haoxuanye Ji, Pengfei Jiang, and Ting Yu; methodology: Pengfei Jiang, Ziyue Wang, and Haoxuanye Ji; software: Zhiliang Chen and Haoxuanye Ji; formal analysis: Pengfei Jiang, Zhiliang Chen, and Haoxuanye Ji; investigation: Pengfei Jiang and Wei Zhang; writing—original draft preparation: Pengfei Jiang, Haoxuanye Ji, and Zhiliang Chen; writing—review and editing: Pengfei Jiang, Haoxuanye Ji, Ziyue Wang, Zhiliang Chen, Ting Yu, and Wei Zhang; supervision: Ting Yu and Wei Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets ANALYZED for this study can be found in the [CUMT-BelT] [https://github.com/CUMT-AIPR-Lab/CUMT-AIPR-Lab] (accessed on 26 September 2025) and the [Cifar-10] and [Cifar-100] [http://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz] (accessed on 26 September 2025).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Cheng D, Xu J, Kou Q, Zhang H, Han C, Yu B, et al. Lightweight network based on residual information for foreign body classification on coal conveyor belt. J China Coal Soc. 2022;47(3):1361–9. (In Chinese).
2. Wang Y, Guo X, Liu X. Design of visual detection system for large foreign body in belt conveyor. Mech Sci Technol Aerospace Eng. 2021;40(12):1939–43. (In Chinese).

3. He M, Wang P, Jiang H. Recognition of coal and stone based on SVM and texture. Comput Eng Design. 2012;33(3):1117–21. (In Chinese).
4. Zhang Y. Research on gangue identification based on video processing [matser's thesis]. Xuzhou, China: China University of Mining and Technology; 2019. (In Chinese).
5. Chen L, Fu Y, Wei K, Zheng D, Heide F. Instance segmentation in the dark. Int J Comput Vis. 2023;131(8):2198–218. doi:10.1007/s11263-023-01808-8.
6. Pu Y, Apel DB, Szmigiel A, Chen J. Image recognition of coal and coal gangue using a convolutional neural network and transfer learning. Energies. 2019;12(9):1735. doi:10.3390/en12091735.
7. Su L, Cao X, Ma H, Li Y. Research on coal gangue identification by using convolutional neural network. In: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC); 2018 May 25–27; Xi'an, China; 2018. p. 810–4.
8. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
9. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324. doi:10.1109/5.726791.
10. Liu Y, Zhang Y, Wang Y, Hou F, Yuan J, Tian J, et al. A survey of visual transformers. IEEE Trans Neural Netw Learn Syst. 2021;35:7478–98. doi:10.1109/tnnls.2022.3227717.
11. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7794–803.
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 x 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2021.
13. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10012–22.
14. Krizhevsky A. Learning multiple layers of features from tiny images. Toronto, ON, Canada: Unversity of Toronto; 2009. Technical Report TR-2009.
15. Whata A, Dibeco K, Madzima K, Obagbuwa I. Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia. Front Artif Intell. 2024;7:1410841. doi:10.3389/frai.2024.1410841.
16. Amaya-Tejera N, Gamarra M, Vélez JI, Zurek E. A distance-based kernel for classification via support vector machines. Front Artif Intell. 2024;7:1287875. doi:10.3389/frai.2024.1287875.
17. Wu L, Shen C, Van den Hengel A. PersonNet: person re-identification with deep convolutional neural networks. arXiv:1601.07255. 2016.
18. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: ICML'20: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Vienna, Austria. p. 1597–607.
19. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Computer Vision—ECCV 2020 (ECCV 2020). Cham, Switzerland: Springer; 2020. p. 213–29.
20. Li Z, Wang W, Li H, Xie E, Sima C, Lu T, et al. BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Computer Vision—ECCV 2022 (ECCV 2022). Cham, Switzerland: Springer; 2022. p. 1–18.
21. Doersch C, Gupta A, Zisserman A. CrossTransformers: spatially-aware few-shot transfer. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 21981–93.
22. Serikbay A, Bagheri M, Zollanvari A, Phung BT. Ensemble pretrained convolutional neural networks for the classification of insulator surface conditions. Energies. 2024;17(22):5595. doi:10.3390/en17225595.
23. Hafeez MA, Munir A, Ullah H. H-QNN: a hybrid quantum—classical neural network for improved binary image classification. AI. 2024;5(3):1462–81. doi:10.3390/ai5030070.
24. Xu J, Jia D, Lin Z, Zhou T, Wu J, Tang L. PBNet: combining transformer and CNN in passport background texture printing image classification. Electronics. 2024;13(21):4160. doi:10.3390/electronics13214160.

25. Ye Z, Li Z, Li G, Zhao H. Dual-channel deep graph convolutional neural networks. Front Artif Intell. 2024;7:1290491. doi:10.3389/frai.2024.1290491.
26. Rublee E, Rabaud V, Konolige K, Bradski G. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2011 Nov 6–13; Barcelona, Spain. p. 2564–71.
27. Harris CG, Stephens MJ. A combined corner and edge detector. In: Alvey Vision Conference; 1988 Aug 31–Sep 2; Manchester, UK. p. 147–52.
28. Paszke A, Gross S, Chintala S, Chanan G, Yang E, Devito Z, et al. Automatic differentiation in PyTorch. In: Proceedings of Advances in Neural Information Processing Systems (NIPS); 2017 Dec 4–9; Long Beach, CA, USA. p. 1–4.
29. Robbins H, Monro S. A stochastic approximation method. Ann Math Statist. 1951;22:400–7. doi:10.1214/aoms/1177729586.
30. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017; 2017 Jul 21–26; Honolulu, HI, USA. p. 6848–56.
31. Ma Y. Rresearch on deep learning algorithm and key technology of coal gangue image recognition [master's thesis]. Xuzhou, China: China Unversity of Mining and Technology; 2009.
32. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015); 2015 Jun 7–12; Boston, MA, USA. p. 1–9.
33. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
34. Cai H, Li J, Hu M, Gan C, Han S. EfficientViT: lightweight multi-scale attention forhigh-resolution dense prediction. In: 2023 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2023); 2023 Jun 17–24; Vancouver, BC, Canada. p. 10012–22.