

Article

GLD-YOLO: An Ore Instance Segmentation Algorithm for Underground Crushing Scenes

Miao Chen ^{1,2}, Yang Liu ¹, Lei Si ^{1,*} , Xitao Wang ² and Jinheng Gu ¹¹ School of Mechanical and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China² Changsha Institute of Mining Research Co., Ltd., Changsha 410012, China

* Correspondence: lei.si@cumt.edu.cn

Abstract

Accurate ore identification is a prerequisite for automated crushing by robotic arms. However, the complex underground mining environment, along with the inherent characteristics of ore such as complex shapes, subtle textures, and blurred boundaries, severely restricts ore identification performance. To address these issues, this paper proposes an improved instance segmentation model, GLD-YOLO, based on the YOLO11-seg network. First, considering the feature aliasing problem in underground ore segmentation, a collaborative enhancement mechanism of C3k2_GBC and C2PSA_LRSA is introduced. By adopting novel gated convolution and local self-attention, the edge and detail features of the target are effectively enhanced, strengthening the saliency of the ore's effective features. Second, addressing the limitation that deep networks easily lose subtle edge information, a PAFPN_DDFE neck network is constructed. Through the novel DDFE module, shallow features are effectively utilized, and the advantages of spatial and frequency domain information are integrated to achieve noise filtering and edge information enhancement. Then, to further refine the boundaries, the gradient update path is optimized using the Focaler-Shape-IoU loss function, guiding the network to focus on difficult samples and improving the regression accuracy of the ore bounding boxes. Finally, comparative experiments and analyses were conducted between the proposed improved method and the benchmark algorithm by building an ore crushing simulation experimental platform. The results show that the proposed method improves the mAP50 and mAP50-95 metrics by 4.4% and 5.1%, respectively, compared to the benchmark model, and the model detection speed reaches 90 FPS. On the MineralImage5k and Crack-seg datasets, the mAP50 is improved by 1.9% and 2.1%, respectively, verifying the feasibility and generalization of the recognition model.

Keywords: YOLO11-seg; attention mechanism; ore instance segmentation; underground mine crushing scene; feature fusion



Academic Editor: Jiangjian Xiao

Received: 19 May 2026

Revised: 17 June 2026

Accepted: 25 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Due to complex geological conditions and limitations in blasting techniques, large pieces of ore that exceed size limits or are highly irregular in shape are often found in the ore during mining operations, which disrupts normal ore crushing operations. Secondary ore breaking is commonly employed to ensure the feed size complies with the specifications of the crushing equipment [1]. However, owing to harsh environmental conditions and limited visibility, manual operations are unable to deliver continuous and precise striking performance. Moreover, this inefficient and high-risk traditional method not only hampers

production efficiency but also fails to ensure the safety of personnel, making sustainable operations impossible [2]. Therefore, the purpose of this study is to improve the automation level of underground ore crushing systems, and the core prerequisite for achieving this goal is to complete the accurate visual recognition of ore targets in crushing scenarios.

In practical crushing scenarios, ore often exhibits severe crowding, stacking, and obstruction, requiring precise identification of the ore and determination of its exact location [3]. To address this problem, this paper establishes a research topic on deep learning-based underground ore instance segmentation algorithms, aiming to determine the specific location of ore using instance segmentation technology. Instance segmentation, as a technique that deeply combines object detection and semantic segmentation, generates masks that can accurately segment the irregular edge contours of the ore, thereby achieving precise separation of adhering ore [4]. Currently, mainstream deep learning segmentation networks are primarily categorized into single-stage and two-stage instance segmentation networks [5]. Among these, two-stage instance segmentation networks are represented by Mask R-CNN. Yang et al. [6] proposed an improved Mask R-CNN model that addresses the issue of detail loss of ore edges during deconvolution in the original algorithm by introducing a hybrid attention module that combines channel attention and spatial attention mechanisms, thereby significantly improving ore segmentation accuracy. To resolve ore blockages at the feed inlet, Luo et al. [7] utilized a Mask R-CNN-based segmentation model to provide precise visual guidance for silos. This approach effectively mitigated errors caused by shifts in geometric center points due to the orientation and irregular shapes of the ore, thereby providing precise visual guidance for mechanical equipment. Shrivastava et al. [8] constructed a Mask R-CNN model using ResNet50 and Feature Pyramid Network (FPN) as the backbone network. This model achieved the precise contour segmentation of irregular ores and significantly reduced the time required to obtain particle size distribution parameters.

Although two-stage instance segmentation networks achieve high accuracy, their high computational complexity and slow inference speeds make it difficult for them to meet real-time requirements. In contrast, single-stage instance segmentation networks, such as YOLO and YOLACT [9], achieve instance segmentation with a single feature extraction step, significantly improving the real-time performance of detection. Jin et al. [10] proposed an LDAFL-YOLOv11 model for detecting foreign objects on coal mine conveyor belts. By introducing a large-separation kernel attention mechanism and a dynamic activation module to optimize the extraction of key features of foreign objects, and by combining this with an attention scale fusion (ASF) mechanism, the model effectively mitigates the issues of missed and false detections of small objects on conveyor belts caused by mutual occlusion and significant size variations. Qiu et al. [11] proposed a lightweight GS-YOLO-seg model that optimizes feature extraction by introducing GSConv and C3k2-Faster modules, and reduces computational complexity by integrating an efficient segmentation head, thereby achieving efficient sorting of low-grade graphite ore. Lv et al. [12] proposed an instance segmentation model based on an improved YOLACT architecture. This model enhances discriminative key features via an attention mechanism, and integrates loss function optimization with an improved non-maximum suppression algorithm. It effectively boosts the segmentation accuracy of ores in heavily occluded scenarios while reducing the false negative rate. Zhang et al. [13] proposed OreNet, a highly efficient anchor-free ore particle size detection model. They constructed a dual-convolution backbone network that integrates clustered convolutions and heterogeneous convolutions, and employed SPPCSPC-DualConv along with a simplified path aggregation network to optimize feature extraction and fusion. This approach addresses the challenges faced by traditional anchor-based methods in adapting to irregular ore shapes and complex backgrounds. Furthermore, to address visual chal-

allenges like uneven lighting, motion blur, and target occlusion, Zhang et al. [14] adapted the SDSE-YOLO architecture. Their method integrates SCINet with deformable convolutions, leading to a much more robust detection performance in harsh operational environments.

The aforementioned studies provide valuable insights into solving the problem of ore identification in fractured scenarios. Current methods mainly rely on spatial feature extraction and uniform local receptive fields. However, these methods still have the following limitations when processing ore segmentation at underground metal grid screens: (1) The complex overlapping and occlusion of ores and the highly similar textures make it easy for existing feature extraction processes to lead to feature aliasing; (2) Single spatial domain feature extraction methods are difficult to separate inherent high-frequency structural noise and highly irregular ore boundaries, resulting in boundary confusion; (3) Conventional detection networks are difficult to achieve high-precision boundary localization and regression when dealing with difficult samples with severe deformation and dense distribution due to the lack of targeted shape feature constraints.

To overcome the bottleneck of feature extraction under complex working conditions and achieve high-precision perception of ore targets, this paper proposes an instance segmentation algorithm based on GLD-YOLO, building upon YOLO11-seg. The feasibility and superiority of the proposed method are verified by constructing an ore crushing experimental platform targeting typical visual interference features in underground mining. The main contributions are as follows:

- (1) To address challenges such as similar mineral textures, blurred boundaries, and uneven illumination, this paper proposes a lightweight gated bottleneck convolution (GBC) and local region self-attention (LRSA) mechanism. In the GBC module, the innovative dual-branch structure and the synergistic effect of perceptual gating dynamically expand the receptive field, improving the network's capability to capture fine textures and irregular geometric features of minerals. Simultaneously, the LRSA module overcomes the limitations of traditional local window partitioning via the overlapping block strategy, effectively alleviating boundary feature loss. The combination of these two modules effectively suppresses environmental noise while significantly enhancing the integrity and saliency of mineral features under severe occlusion and similar textures.
- (2) Due to the tendency of deep networks to lose subtle features and the high-frequency noise interference induced by metal sieves, and considering the limitations of single-spatial-domain feature extraction, a PAFPN_DDFE neck network was designed. This network not only innovatively introduces shallow information from P2 and P3, but also proposes an innovative DDFE module, which effectively decouples back-ground noise in the frequency domain from the complex edges of the ore by introducing wavelet transform. Furthermore, it incorporates a deformable large-kernel convolution to achieve dynamic adaptive fitting of irregular ore deformations in the spatial domain. This spatial-frequency dual-domain complementary mechanism significantly enhances the network's ability to perceive and distinguish the edges of complex, densely adhered ores.
- (3) In underground ore segmentation scenarios, complex samples with high overlap and blurred contours account for a large proportion, which can easily lead to degradation of model detection accuracy. This paper constructs an attention-guided mechanism for hard-to-regress bounding boxes by reshaping the optimization gradient of the model, so that the network pays more attention to the target boundaries that are difficult to fit accurately during the training process. This method accelerates the overall convergence speed of the network, and effectively improves the bounding

box regression accuracy and generalization performance of the model in multi-scale stacked ore scenarios.

The rest of the paper is structured as follows. Section 2 details the theoretical basis and architecture of the proposed GLD-YOLO model, including its enhanced feature extraction module, PAFPN_DDFE neck network and improved loss function. Section 3 describes the experimental setup and evaluation metrics, and presents a comprehensive analysis of the model's segmentation performance, ablation results and cross-scene generalization capability. Finally, this paper concludes with a summary of the work, along with discussions on its limitations and prospects for future research.

2. Materials and Methods

This section introduces the theoretical foundation and structural design of the proposed GLD-YOLO model, aiming to improve inspection accuracy by addressing issues such as feature aliasing. The chapter is structured as follows: Section 2.1 introduces the overall network architecture; subsequently, Sections 2.2 and 2.3 elaborate on the improved feature extraction mechanism and the dual-domain PAFPN_DDFE neck network, respectively; finally, Section 2.4 explains the loss function specifically optimized for dense ore bounding box regression.

2.1. GLD-YOLO

To achieve accurate ore identification in underground crushing scenarios, this paper proposes an ore instance segmentation method based on GLD-YOLO, the network structure of which is shown in Figure 1.

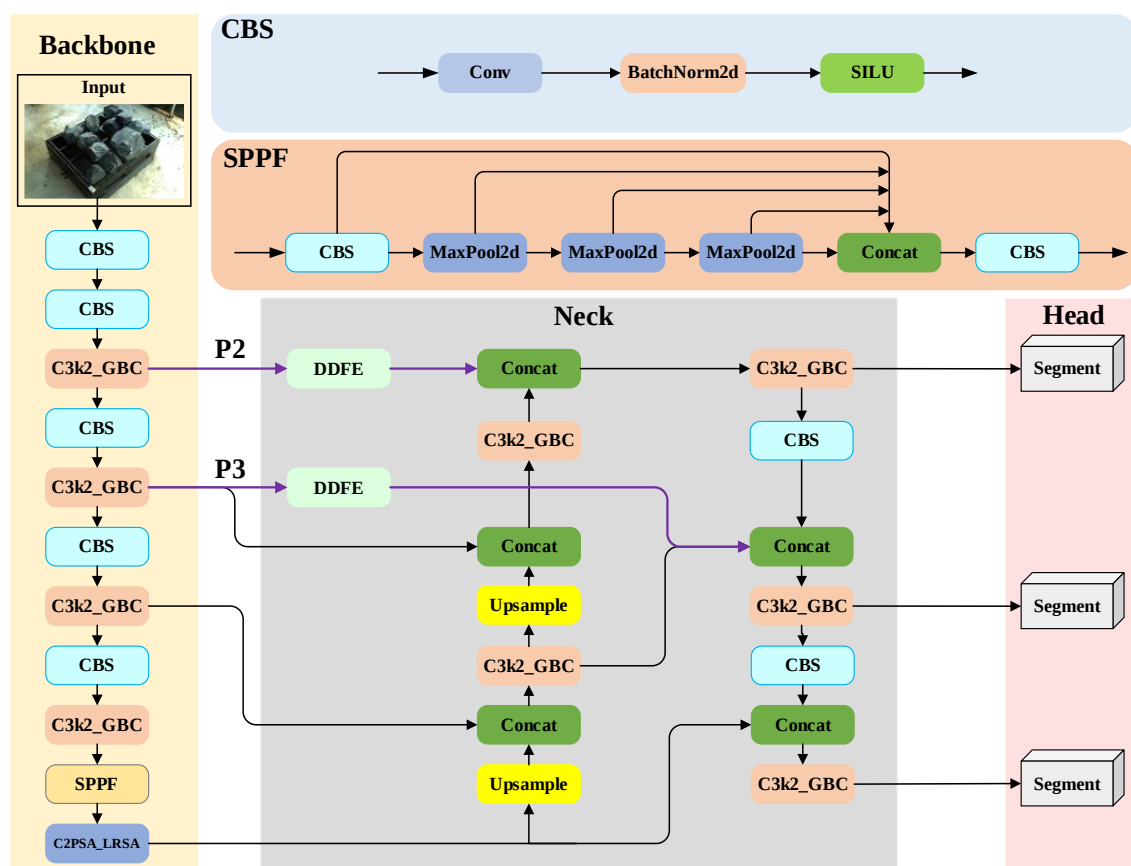


Figure 1. GLD-YOLO Network Architecture.

The proposed network is designed based on the YOLO11-seg framework [15] and consists of three core components: a feature-extracting backbone, a dual-domain feature fusion neck, and a multi-scale segmentation head. Image processing commences in the backbone, transitioning from initial shallow extraction at the Conv-BatchNorm-SiLU (CBS) module to the advanced C3k2_GBC unit formulated in this study. The C3k2_GBC deploys a lightweight dual-branch layout with perceptual gating to expand the receptive field, thereby enhancing the morphological and textural representation of the ores. This effectively mitigates feature aliasing caused by visual similarities while strictly maintaining computational economy. Subsequently, deep features synthesized by the SPPF module are directed into the enhanced C2PSA_LRSA aggregator. The LRSA module adopts an overlapping patch strategy, which effectively mitigates the problems associated with localized window splitting. These methods are designed to enhance the saliency of discriminative ore features under heavy shading conditions.

Subsequently, following multi-scale extraction, the features proceed to the PAFPN_DDFE neck for cross-layer aggregation. Because high-frequency noise heavily corrupts the shallow P2 and P3 layers, the custom DDFE module is deployed explicitly as a filtering gateway. It leverages frequency-domain wavelet transforms to decouple actual ore boundaries from the noise, while simultaneously employing deformable large-kernel convolutions to adapt to morphological variations. These purified representations are then concatenated with upsampled deep features and refined by the C3k2_GBC unit. This complementary spatial-frequency strategy effectively preserves vital shallow features, which contributes to better boundary localization accuracy for dense ores. Ultimately, the fused multi-scale data is routed to three distinct Segment heads for pixel-level prediction. To optimize this stage, the Focaler-Shape-IoU loss is implemented during training. By imposing geometric shape constraints, this formulation forces the model to prioritize difficult, highly overlapping samples with vague contours. This approach effectively alleviates the problem of unbalanced sample distribution, accelerates the convergence speed, and further breaks through the bottleneck of bounding box regression accuracy in densely stacked scenarios.

2.2. Feature Extraction and Aggregation Network Structure

In the context of underground grid screening, where ore is densely stacked, has blurred boundaries, similar textures, and is subject to strong background noise interference, baseline models struggle to balance the ability to represent the fine features of the ore with the ability to resist interference in complex scenarios. Therefore, this paper improves the feature extraction and feature aggregation stages in the network by proposing the C3k2_GBC high-efficiency feature extraction module and the C2PSA_LRSA feature aggregation module, thereby optimizing the model's ability to perceive ore targets.

When faced with tasks involving segmenting stacked and occluded ore instances, traditional convolutional units within the standard C3k2 bottleneck structure typically apply uniform weights across all spatial locations within their local receptive field. In underground crushing scenarios, adjacent ores often have highly similar surface textures and lack significant boundary gradients. Since standard convolutions cannot dynamically adjust their spatial attention based on image content, they indiscriminately aggregate feature representations from closely connected neighboring objects. This inherent spatial uniformity computationally mis-merges the boundary responses of neighboring instances, frequently leading to feature aliasing between clustered objects. Therefore, to break this spatial uniformity, this paper introduces gated bottleneck convolution (GBC) into the C3k2 module. Structurally, the GBC replaces standard single-path convolutions with a dual-branch architecture, a deep semantic branch to extract features and a structural gating branch to generate spatial attention weights. These branches interact to highlight

essential structural details while suppressing irrelevant background noise. The C3k2_GBC network structure is shown in Figure 2. Leveraging the inherent lightweight and low-rank characteristics as well as the adaptive gating mechanism of the GBC module, the proposed method enables precise enhancement of critical ore geometric boundary features while effectively mitigating feature confusion in densely stacked occlusion areas.

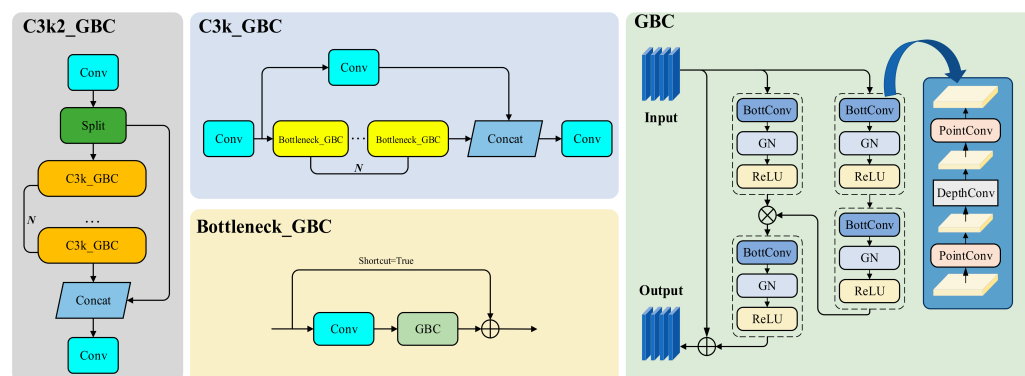


Figure 2. C3k2_GBC Network Architecture.

This module not only possesses the advantage of the rich gradient flow inherent in the Cross Stage Partial (CSP) architecture but also deeply integrates the gating enhancement characteristics of GBC. First, after passing through a 1×1 convolutional layer, the input features are divided into two branches: one branch retains shallow feature information through a cross-stage direct connection, facilitating the rapid backpropagation of gradients; the other branch enters the main feature extraction path, which consists of multiple stacked C3k_GBC units. The interior of the C3k_GBC unit adopts the Bottleneck_GBC structure. This structure replaces the core operator with the GBC module and promotes the cross-layer propagation of gradients through residual connections. Finally, the deep abstract features output by the main feature extraction path and the shallow features retained by the cross-stage direct connection branch are concatenated along the channel dimension.

The GBC module, as the core component of C3k2_GBC, is mainly constructed from the BottConv basic unit. BottConv adopts an inverse residual design from pointwise convolution [16] to depthwise convolution and then to pointwise convolution, using 1×1 convolution to adjust the channel dimension and combining it with depthwise convolution to extract spatial features. In order to adapt to small batch training and enhance the statistical stability of features, group normalization is introduced to replace the traditional batch normalization.

At the feature interaction level, the GBC module replaces the conventional single feed-forward pathway with a dual-branch structure-aware gating mechanism, dividing the input features into a deep semantic branch and a structural gating branch. Specifically, the deep semantic branch employs two cascaded BottConv modules coupled with ReLU activations to perform deep non-linear transformations, thereby extracting high-dimensional discriminative features. The structural gating branch employs a single BottConv module to produce spatial attention weights, focusing on key structural details like edges. The outputs from both branches are then combined through element-wise multiplication, allowing the gating signal to highlight essential features while filtering out irrelevant noise. Finally, the fused representation passes through another BottConv module and is added to the original input via a residual connection. The complete operation of the GBC module is expressed as Equation (1):

$$Y = \phi_4(\phi_2(\phi_1(X)) \odot \phi_3(X)) + X \quad (1)$$

where $\phi_i(\cdot)$ represents the BottConv transformation block consisting of bottleneck convolution, Group Normalization, and an activation function; $\odot$ denotes element-wise multiplication; and $+$ indicates the residual connection.

On the other hand, during the transmission of deep features, uneven underground illumination and high-frequency noise can easily lead to a severe loss of detailed information [17]. Under such extreme contrast changes, traditional global feature extraction methods are often affected by severe illumination deviations, and it is difficult to maintain the effectiveness of features by relying solely on C3k2_GBC to extract features. To this end, this paper introduces C2PSA_LRSA at the end of the backbone network, which is based on the LRSA mechanism introduced by the C2PSA module. Structurally, this module divides the input features into overlapping patches to perform multi-head self-attention, and it integrates a Convolutional Feed-Forward Network. Through its overlapping block strategy and this dual-path refinement, this mechanism overcomes the limitations of traditional local window segmentation, enhances the edge features of the ore body, and further strengthens the saliency expression of effective ore body features. The network structure of C2PSA_LRSA is shown in Figure 3.

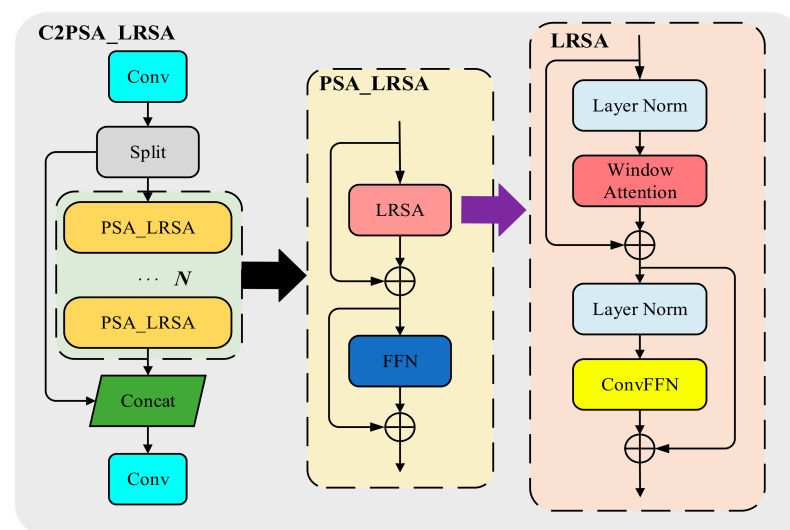


Figure 3. C2PSA_LRSA Network Architecture.

First, C2PSA_LRSA follows the split-flow concept of the CSP architecture, dividing the input feature map X into two branches along the channel dimension: the main branch undergoes channel dimensionality reduction via a 1×1 convolution before being fed into the stacked PSA_LRSA units, while the other branch retains the original feature information. Finally, feature fusion is achieved through concatenation and convolution operations. Within the PSA_LRSA unit, the LRSA core integrates a multi-head self-attention mechanism and adopts a synergistic architecture of overlapping patching and a Convolutional Feed-Forward Network (ConvFFN), employing a residual connection that integrates the identity mapping with features processed through Layer Normalization and the ConvFFN. The oversampled overlapping design constructs an information redundancy mechanism for the features, ensuring that even when local regions are affected by factors such as sudden illumination changes, key semantic cues can still be preserved and reused through the overlapping areas of adjacent windows.

The specific processing flow is as follows: During the forward propagation of the LRSA module, the input feature map first undergoes a patching operation via sliding windows, cropping the image into a series of overlapping local region tensors X_p . Subsequently, the local region tensors are fed into the Window Attention module, where a local-window-

based multi-head self-attention mechanism is employed for deep feature interaction. By mapping local features into multiple independent subspaces, this mechanism captures long-range dependencies from different semantic perspectives. Furthermore, by calculating the affinity matrix between pixels, it dynamically allocates attention weights, enabling the network to focus on high-confidence regions of the ore targets while suppressing dust or background noise. Afterward, the processed local regions are restored to the original image dimensions through a local region reconstruction operation. The core computational formulas of this process are expressed as Equations (2) and (3).

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

$$X'_p = \text{LRSA}(X_p) = \text{LRR}(\text{Attention}(X_p) + X_p) \quad (3)$$

where Q , K , and V are respectively generated from the local region tensor X_p through linear projection; d_k is the scaling factor; LRR denotes the average pooling of overlapping regions and local region reconstruction; and Attention represents the local-window-based multi-head self-attention computation mechanism.

To compensate for the deficiency of the self-attention mechanism in local inductive bias, a ConvFFN unit is appended at the end of the LRSA module. After passing through Layer Normalization, the input features enter the ConvFFN and undergo residual fusion with the features prior to entering this sub-layer. The ConvFFN unit embeds a depthwise separable convolution [18] between two fully connected layers, utilizing the depthwise separable convolution to capture pixel-level local spatial correlations in a lightweight manner. The specific mathematical expression is shown in Equation (4).

$$Y = X + \text{ConvFFN}(\text{LN}(X)) \quad (4)$$

where X is the input feature, and $\text{LN}(\cdot)$ denotes Layer Normalization; $+$ represents the residual connection that fuses the features from the convolutional branch and the linear branch. The internal computational logic of the ConvFFN is as follows:

$$\text{ConvFFN}(x) = L_2(\text{GELU}(\text{DW}(L_1(x)))) \quad (5)$$

where $L_{1,2}$ denote the fully connected layers, GELU represents the Gaussian Error Linear Unit activation function, and $\text{DW}(\cdot)$ indicates the depthwise separable convolution.

2.3. PAFPN_DDFE Neck Network Structure

In practical mining scenarios, ores typically exhibit diverse shapes, irregular boundaries, and highly similar textures. Furthermore, at the feeding inlet, these materials are often densely stacked and heavily occluded. The standard YOLO11-seg architecture utilizes a conventional Path Aggregation Feature Pyramid Network (PAFPN). In underground crushing scenarios, on the one hand, mechanical crushing causes highly irregular multi-scale geometric deformation of the ore; on the other hand, metal screens and the environment generate high-frequency noise, which severely interferes with ordinary shallow features. Single spatial domain or frequency domain extraction methods are ineffective.

For solving the above problems, this paper presents a new PAFPN_DDFE neck network. In this architecture, the newly designed Dual-Domain Feature Extraction (DDFE) module functions as a filtering gateway within the Neck network to effectively utilize shallow features. Structurally, this gateway combines spatial-domain deformable convolutions with frequency-domain wavelet transforms to effectively suppress high-frequency background noise and enhance the feature representation. In this network, for the P2

feature with the highest spatial resolution, it is input as a whole into the DDFE for dual-domain denoising and feature refinement, and then connected to the top aggregation node of the FPN. Meanwhile, for the P3 feature, a dual-branch routing mechanism is adopted. Its original feature directly participates in the top-down semantic fusion of FPN, while the other branch of the feature is filtered for noise through DDFE and directly injected as a cross-layer bypass signal into the feature enhancement path of the path aggregation network (PAN), thereby completing feature enhancement and fusion.

The structure of the DDFE module is shown in Figure 4a. The core design of the DDFE module is to utilize a dual-domain complementary mechanism to address the feature aliasing problem in complex ores. Spatial features are acquired through Edge-Aware Deformable Large Kernel Convolution (EDLKC), while the frequency-domain information of shallow features is extracted using Wavelet Transform Convolution (WTConv). Subsequently, Cross-Attention is employed to achieve an effective integration of spatial geometric information and frequency-domain texture information, followed by lossless downsampling and feature refinement via the Space-to-Depth (SPD) operation.

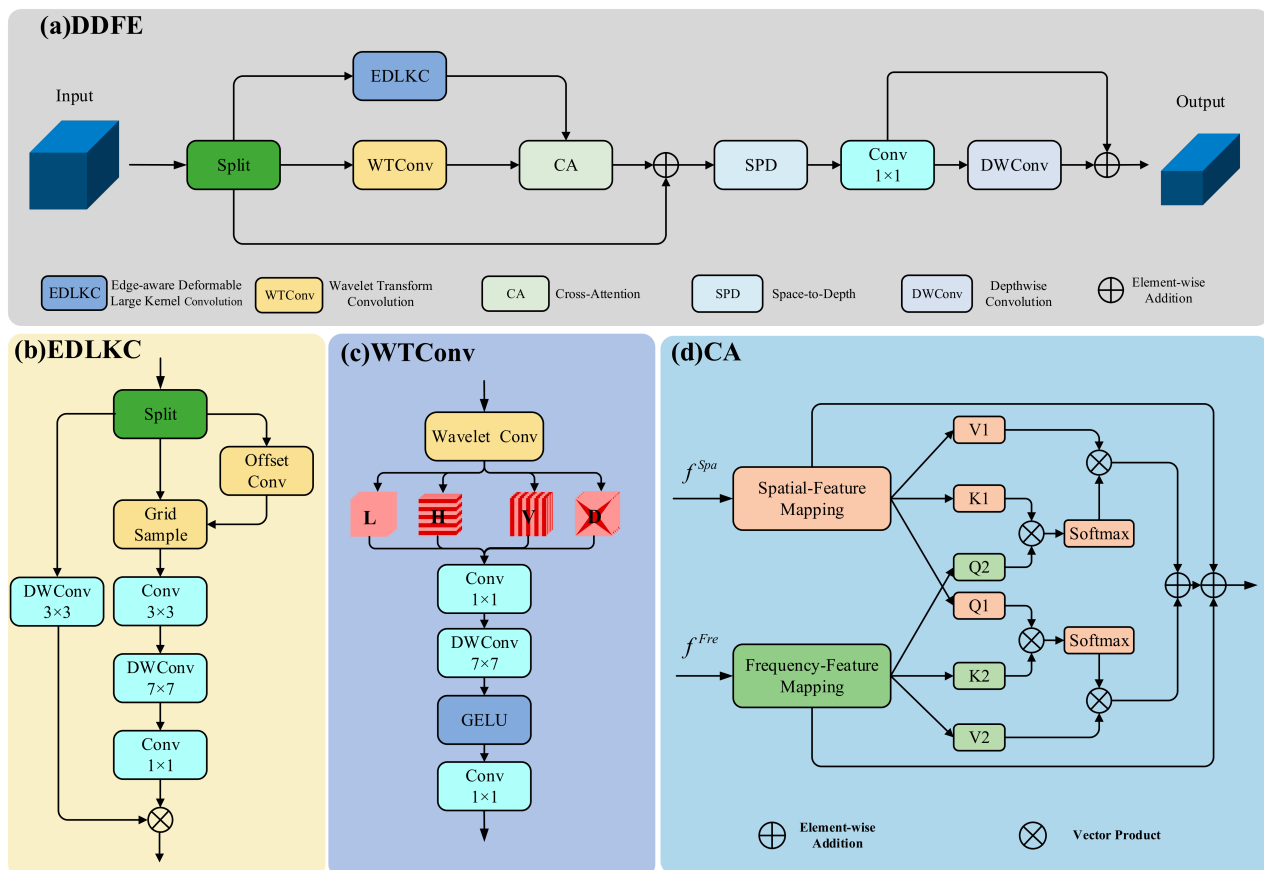


Figure 4. DDFE Network Architecture.

Considering the highly irregular shapes and severe stacking of ore, this paper draws on the adaptive sampling mechanisms of Large Kernel Attention (LKA) [19] and Deformable Convolutional Networks (DCN) [20] to design Edge-Aware Deformable Large Kernel Convolution (EDLKC), which aims to balance the modeling of long-range dependencies with the ability to adapt to local geometric deformations. Its structure is shown in Figure 4b. EDLKC adopts a dual-path parallel processing architecture: first, the right path uses offset convolutions (Offset Conv) to capture the geometric distortion features of the ore, generating adaptive positional offsets; subsequently, a feature resampling mechanism is employed to realign the feature maps based on the generated offsets. The aligned features are then

sequentially processed by a 3×3 convolution to capture local details, a 7×7 depthwise separable convolution to establish long-range dependencies, and a 1×1 convolution for channel interaction. This design allows the model to simultaneously extract fine textures and global morphology from the aligned feature map with low computational overhead. Meanwhile, the left branch employs a 3×3 depthwise separable convolution to form a local detail-sensing sub-network. Its output is element-wise multiplied with the right-branch features, adaptively suppressing background noise and enhancing the ore's edge contours.

Since individual ore particles exhibit highly similar textures, the detection task relies heavily on edge features, while the background metal grid sieve introduces intense periodic high-frequency noise interference. Therefore, this paper employs Wavelet Transform Convolution (WTConv). Taking advantage of the spatial-frequency analysis of wavelet transform, this module decomposes image features into distinct frequency bands. It then uses the differences in these frequency components to decouple structural information from periodic noise, as shown in Figure 4c.

WTConv first performs a lossless decomposition of the input features using the Discrete Wavelet Transform (DWT) to obtain low-frequency (L) and high-frequency components in the horizontal (H), vertical (V), and diagonal (D) directions. These components are then concatenated along the channel dimension. To address interference caused by uneven illumination (primarily in the low-frequency band) and to separate ore edge details from high-frequency grid noise, the module employs cascaded 1×1 and 7×7 depthwise separable convolutions to recalibrate the concatenated features in the frequency domain. This approach suppresses environmental noise while selectively enhancing the high-frequency responses of the sharp ore boundaries. After GELU activation, the calibrated features from each frequency band are concatenated again, followed by a 1×1 convolution for channel reduction and feature aggregation.

DDFE employs a cross-attention mechanism (CA) to deeply fuse spatial and frequency-domain features, as illustrated in Figure 4d. This mechanism aims to establish an interactive mapping between the spatial features f^{Spa} output by EDLKC and the frequency-domain enhanced features f^{Fre} output by WTConv. Specifically, the input features first pass through spatial and frequency-domain mapping layers to generate the corresponding queries (Q), keys (K), and values (V). The spatial branch generates $\{Q_1, K_1, V_1\}$, while the frequency-domain branch generates $\{Q_2, K_2, V_2\}$. Next, the module executes two parallel cross-domain interaction paths. The first reorganizes spatial features using the frequency-domain texture distribution as a guide. The second uses the ore's spatial location to retrieve high-confidence edge details from the frequency domain, compensating for missing texture information in the spatial domain. Finally, the weighted features from both paths are aggregated and added to the original input via a residual connection. The core fusion process of the CA module is formulated as:

$$f_{out} = S\left(\frac{Q_2 K_1^T}{\sqrt{d_k}}\right) V_1 + S\left(\frac{Q_1 K_2^T}{\sqrt{d_k}}\right) V_2 + f^{Spa} + f^{Fre} \quad (6)$$

where $S(\cdot)$ denotes the Softmax activation function, and f_{out} is the final fused output feature of the cross-attention mechanism; d_k is the scaling factor for the channel dimension of the key vectors; the first two terms on the right side of the equation represent the frequency-guided spatial attention feature and the spatial-guided frequency attention feature, respectively, while the last two terms are the original features establishing the residual connection.

2.4. Loss Function Design

In underground ore instance segmentation, training samples often suffer from a severe imbalance between easy and hard examples. This is primarily caused by uneven illumination and the diverse morphologies of the ores. Specifically, the background and highly salient ores constitute the vast majority of easy samples. In contrast, irregularly shaped ores with dense stacking and blurred boundaries form hard samples that are difficult to regress. Standard bounding box regression losses are typically dominated by these easy samples, causing the model to struggle with precise localization for densely clustered targets [21]. To address this limitation, we adopt Focaler-Shape-IoU as the bounding box regression loss function. By integrating a sample-focusing mechanism with geometric shape constraints, this method aims to improve the model's localization accuracy and overall detection performance.

For dense ore detection, Focaler-IoU is introduced to address the imbalance between easy and hard samples. This mechanism modifies the standard Intersection over Union (IoU) metric through linear interval mapping, allowing for the dynamic reallocation of sample weights during regression. Specifically, Focaler-IoU lowers the loss weight of easily detected isolated ores, guiding the network to concentrate on severely overlapped targets with blurred boundaries. The reformulated IoU is expressed in Equation (7):

$$IoU^{focaler} = \begin{cases} 0, & IoU < d \\ \frac{IoU-d}{u-d}, & d \leq IoU \leq u \\ 1, & IoU > u \end{cases} \quad (7)$$

where $IoU^{focaler}$ is the reconstructed IoU, and $[d, u] \in [0, 1]$ are the hyperparameters of the linear mapping interval. By adjusting the thresholds of d and u , $IoU^{focaler}$ can be tailored to focus on different regression samples. In this paper, to effectively address the severe occlusion in dense ore scenes, the parameters are empirically set to $d = 0$ and $u = 0.95$. Specifically, this configuration suppresses the gradient contribution of easily localized targets ($IoU > 0.95$) while fully retaining that of hard samples, thereby forcing the network to focus on densely stacked and blurred ores.

Moreover, since ores exhibit irregular shapes and multi-scale variations, an IoU loss based solely on spatial intersection is often insufficient to characterize the alignment between predicted and ground-truth boxes. To mitigate this limitation, we incorporate the shape-aware penalty term from Shape-IoU. Unlike conventional distance metrics, this formulation directly considers the geometric properties and scale of the bounding boxes. By introducing the shape-adaptive weight Ω^{shape} and the shape-aware distance loss $distance^{shape}$, the regression process becomes more sensitive to the aspect ratios and irregular contour changes in the ores. The core formula of Shape-IoU is shown in Equation (8):

$$L_{Shape-IoU} = 1 - IoU + distance^{shape} + 0.5 \times \Omega^{shape} \quad (8)$$

Ultimately, to simultaneously account for the balanced learning of dense samples and the geometric precision of irregular targets, this paper constructs the fused loss function Focaler-Shape-IoU. By utilizing $IoU^{focaler}$ to replace the standard IoU, this loss function significantly enhances the model's robustness in addressing the issues of dense stacking and adhering occlusion, while maintaining high sensitivity to the geometric shape features of the ores. Its final definition is shown in Equation (9):

$$L_{Focaler-Shape-IoU} = 1 - IoU^{focaler} + distance^{shape} + 0.5 \times \Omega^{shape} \quad (9)$$

It is noteworthy that although Focaler-Shape-IoU operates exclusively on the bounding box regression loss, it substantially enhances the instance segmentation performance of GLD-YOLO. This is because in the segmentation paradigm employed in this work, the prediction quality of the mask branch is intrinsically coupled to the spatial localization accuracy of the detection branch. For densely stacked ore targets in underground mines, Focaler-Shape-IoU, by enhancing the focus on hard samples and imposing stricter geometric constraints, drives the network to generate tighter and more precise bounding boxes. This refined spatial prior helps to mitigate feature interference from adjacent targets, constrains the mask generation range, and contributes to reducing the false positive rate at edge pixels. Therefore, optimizing bounding box regression capability directly translates to improved segmentation mask quality, enabling the model to produce instance masks with sharper edges and superior target alignment when processing densely stacked ore targets with blurred boundaries.

3. Results and Discussion

This section aims to comprehensively evaluate the segmentation performance of the GLD-YOLO model. Sections 3.1–3.4 first present an overview of the experimental platform, dataset, and evaluation metrics. Sections 3.5–3.8 then demonstrate the model's superiority through comparisons of different models, ablation experiments, and module-specific comparisons. Finally, Section 3.9 evaluates the model's generalization ability in different scenarios employing publicly available datasets.

3.1. Experimental Platform Setup

To verify the adaptability and identification effectiveness of the proposed instance segmentation network model for complex underground crushing conditions, an ore-crushing simulation experimental platform and site were established, as shown in Figure 5. The experimental platform primarily consists of industrial cameras, mobile floodlight, a crushing mechanical arm, a metal grating, and experimental test blocks.

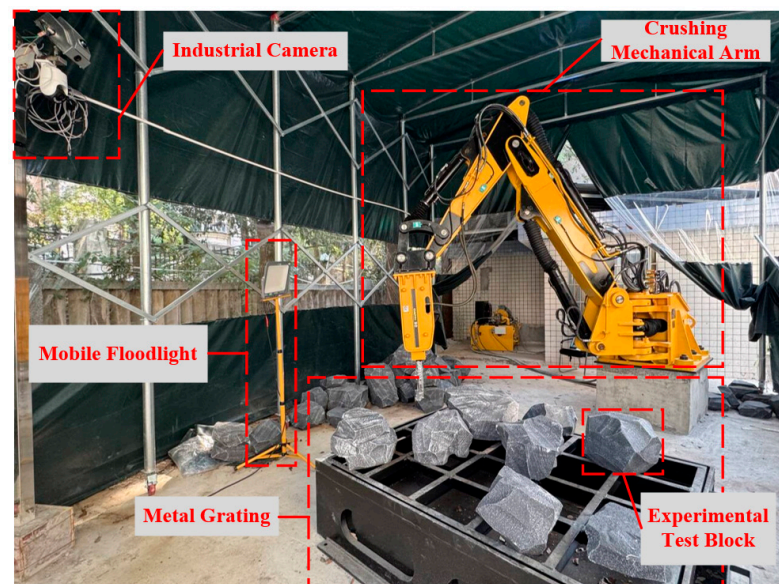


Figure 5. Ore Crushing Simulation Experimental Platform.

In this experimental platform, the crushing equipment is a multi-degree-of-freedom hydraulic robotic arm equipped with a hydraulic breaker, which is used to simulate the actual crushing operation of large ore blocks in underground mines. A monocular vision system is adopted to acquire ore image data. The camera model is has a resolution

of 2048×1200 pixels, equipped with a 1/1.7-inch CMOS sensor and supporting GigE transmission. The lighting device is a 120 W high-power LED mobile floodlight, which allows height and angle adjustment. It is designed to simulate the uneven illumination and shadow environment formed by point light sources in underground mine working conditions. A high-strength metal grid sieve with a size of $500 \text{ mm} \times 500 \text{ mm}$ is arranged in the experimental scene. The recognition targets are irregular experimental test blocks scattered on the grid sieve, which simulate the complex working conditions of mutual extrusion, occlusion and edge blurring of ores in actual underground crushing scenarios.

3.2. Dataset

This paper collected 1900 ore images through ore crushing experiments. 200 images of the same ore category were selected from publicly available ore datasets for augmentation and annotation to create a dedicated dataset. The ratio of normal lighting images to non-uniform lighting images was 7:3, and stacked scenes accounted for 55%. The total number of annotated instances was approximately 21,420, with an average of 10.2 ore elements per image. Finally, the ore dataset was randomly divided into training, validation, and test sets in an 8:1:1 ratio, resulting in 1680 training images, 210 validation images, and 210 test images. These were used to evaluate the performance of the ore instance segmentation model. Sample examples of some ores are shown in Figure 6.

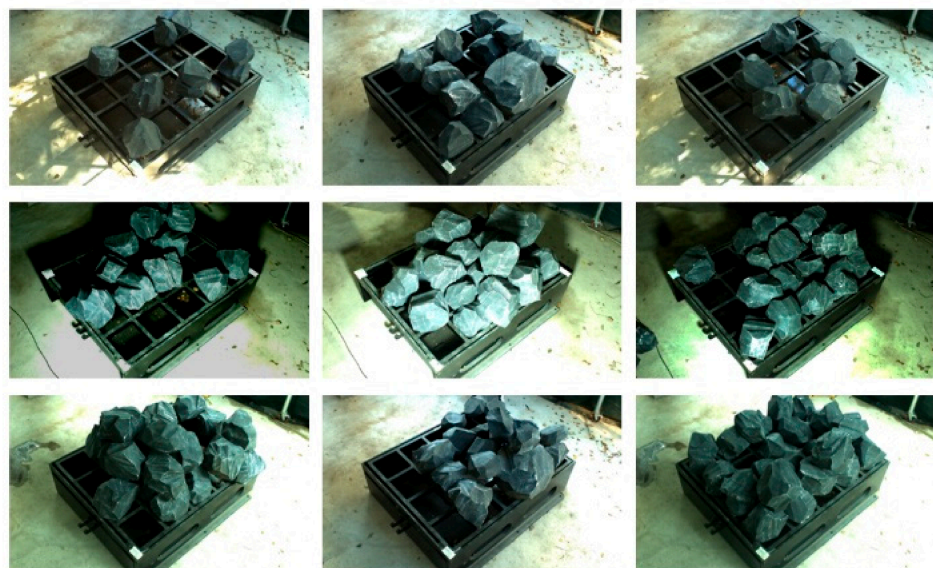


Figure 6. Example of self-constructed ore dataset.

In addition, this paper further verifies the generalization ability of the model by introducing the MineraImage5k [22] and Crack-seg [23] datasets. The MineraImage5k dataset comprises 5341 high-resolution images with over 30,000 instance annotations, characterized by extremely high inter-class similarity and complex texture features. The diverse colors, shapes, and crystal structures of the objects make this dataset ideal for validating the model's performance in multi-texture ore recognition and fine-grained classification tasks. Additionally, Crack-seg is a public dataset focused on structural defect detection, containing 4717 crack images collected from roads, bridges, and walls. Targets in this dataset possess extreme aspect ratios and irregular topologies. Moreover, since crack edges often blend with background noise, Crack-seg acts as an ideal benchmark to evaluate the model's capacity for irregular object extraction and blurred edge segmentation.

3.3. Experimental Evaluation Metrics

The performance of the proposed model is evaluated using Precision (P), Recall (R), mean Average Precision (mAP), the number of parameters (Params), and GFLOPs.

Precision measures the proportion of correctly predicted positive samples out of all predicted positives. It represents the fraction of genuinely correct ore detections among all instances identified as ore. Here, TP (True Positive) and FP (False Positive) denote the number of correctly and incorrectly detected targets, respectively.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

Recall measures the proportion of actual positive instances that are correctly detected by the algorithm. A higher recall indicates that fewer real ore targets are missed. FN (False Negative) represents the number of ground-truth targets that the model fails to detect.

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

Average Precision (AP) measures the detection performance for a single class, while mAP represents the mean AP across all classes. Specifically, mAP₅₀ denotes the mAP at an IoU threshold of 0.5, and mAP_{50–95} represents the average mAP evaluated across varying IoU thresholds from 0.5 to 0.95. For instance segmentation, a higher mAP score strictly indicates superior localization and segmentation accuracy.

$$AP = \int_0^1 P(R) dR \quad (12)$$

$$mAP = \frac{1}{n} \sum_{i=0}^n AP_i \quad (13)$$

Given that the core task of this study is ore instance segmentation, the mAP₅₀ and mAP_{50–95} metrics presented in subsequent experiments all refer to Mask Average Precision.

Finally, given the real-time requirements of the robotic arm visual guidance system in underground crushing operations, Frames Per Second (FPS) is employed to evaluate the model's inference speed. FPS denotes the number of images processed by the model per second. During practical testing, the FPS calculation accounts for the total latency, including image preprocessing, network forward inference, and Non-Maximum Suppression (NMS) post-processing.

3.4. Experimental Environment

All experiments are conducted on a Windows 11 system equipped with an Intel Core i5-14600KF CPU and an NVIDIA GeForce RTX 5070 GPU. The models are implemented in Python 3.11.11 using the PyTorch 2.7.0 framework with CUDA 12.8. During training, to strictly ensure the absolute fairness of the comparative evaluation among various heterogeneous architectures, all networks are trained from scratch without any pre-trained weights, and optimized using the Stochastic Gradient Descent (SGD) algorithm. Detailed experimental hyperparameters are summarized in Table 1.

The specific calculation process for all experimental metrics is standardized as follows. First, the number of parameters and computational complexity (GFLOPs) of all models are derived using static analysis based on a standard input resolution of 640 × 640 using the PyTorch profile tool. Second, for the accuracy metrics (P, R, mAP), the model weights that achieve the highest accuracy on the validation set during 300 training epochs are extracted and deployed on an independent test set. The final metrics are calculated by comparing

the predicted mask with the true label. Finally, to measure the inference speed (FPS) in Table 2, the test set images are input one by one with a batch size of 1, and the average total time, including preprocessing, forward inference, and NMS post-processing, is recorded to calculate the FPS value in the experimental environment.

Table 1. Experimental parameters.

Parameter Name	Parameter Setting
Epochs	300
Batch Size	16
Image Size	640 × 640
Model Size	depth: 0.25, width: 0.25
Initial Learning Rate	0.01
Learning Rate Momentum	0.937
Weight Decay	0.0005
LR Scheduler	Cosine Annealing

Table 2. Comparison of Instance Segmentation Models.

Model	P (%)	R (%)	mAP ₅₀	mAP _{50–95}	Params (M)	GFLOPs	FPS
Mask-RCNN	88.1	87.3	88.4	74.3	46.4	239.4	18
YOLACT	88.4	87.7	88.9	74.9	31.3	61.2	48
SOLOv2 [24]	88.9	88.4	91.5	75.3	41.6	168.2	24
YOLOv8n-seg	90.7	90.4	92.4	76.7	2.77	11.2	112
YOLOv9t-seg	90.6	90.2	92.6	76.6	3.13	54.4	55
YOLOv10n-seg	90.4	89.9	91.9	76.2	2.85	11.8	115
YOLO11n-seg	91.1	90.9	93.7	77.2	2.84	10.4	120
YOLO12n-seg	90.9	90.5	93.4	77.1	2.79	9.9	125
YOLO13n-seg	91.1	90.4	93.1	76.9	2.71	10.3	122
FastInst	89.6	89.2	92.1	75.8	36.5	82.4	34
RF-DETR	92.4	91.8	94.5	78.6	33.6	75.5	38
GLD-YOLO	95.9	95.6	98.1	82.3	3.52	16.1	90

3.5. Performance Comparison of Instance Segmentation Models

To evaluate the proposed GLD-YOLO in underground crushing scenarios, ten mainstream instance segmentation models are selected as baselines for comparison. These include Mask-RCNN, YOLACT, SOLOv2, five YOLO variants (YOLOv8n-seg, YOLOv9t-seg, YOLOv10n-seg, YOLO12n-seg, YOLO13n-seg), and two Transformer-based models (FastInst and RF-DETR). To ensure a fair comparison, all models are trained on the same dataset. The quantitative results, measured in terms of P, R, mAP₅₀, mAP_{50–95}, GFLOPs, Params, and FPS, are summarized in Table 2.

The experimental results demonstrate that GLD-YOLO achieves the best overall performance. Compared to Mask-RCNN, YOLACT, and SOLOv2, GLD-YOLO yields improvements of 7.0–7.8% in Precision, 7.2–8.3% in Recall, 6.6–9.7% in mAP₅₀, and 7.0–8.0% in mAP_{50–95}. Furthermore, it requires significantly fewer parameters and lower computational overhead, while exhibiting a substantial advantage in inference speed.

When compared to the YOLO series models, GLD-YOLO achieves the highest scores across Precision, Recall, mAP₅₀, and mAP_{50–95}. Specifically, relative to the baseline YOLO11n-seg, the proposed algorithm delivers substantial accuracy gains, boosting Precision by 4.8%, Recall by 4.7%, mAP₅₀ by 4.4%, and mAP_{50–95} by 5.1%. Although the inference speed experiences a slight decrease compared to the 120 FPS of the baseline model, GLD-YOLO still maintains a robust rate of 90 FPS, which to some extent meets the real-time processing requirements of industrial systems.

Furthermore, when compared to FastInst and RF-DETR, two recent Transformer-based instance segmentation networks, GLD-YOLO surpasses RF-DETR by 3.5%, 3.6%, and 3.7% in Precision (P), mAP₅₀, and mAP₅₀₋₉₅, respectively. It also outperforms FastInst by 6.3%, 6.0%, and 6.5% across the same three metrics. Meanwhile, GLD-YOLO maintains a noticeably smaller parameter size and lower computational complexity, achieving an effective balance between lightweight design and high precision for complex edge segmentation tasks.

In summary, the improved model proposed in this study delivers excellent accuracy. While maintaining real-time inference capability, it achieves an effective balance between detection accuracy and inference speed under limited computational resources.

3.6. Ablation Experiments

This paper selects YOLO11n-seg as the baseline architecture to explore the specific contributions of each improved module to the model performance, and conducts ablation experiments while keeping the experimental environment completely consistent. As shown in Table 3, the “√” symbol indicates that the model integrates the corresponding improvement strategy. The table details the performance metrics and comparison results of each experimental scheme.

Table 3. Ablation study results.

C3k2_GBC	C2PSA_LRSA	PAFPN_DDFE	Focaler-Shape-IoU	P/%	R/%	mAP ₅₀	mAP ₅₀₋₉₅	Params/M	GFLOPs
				91.1	90.9	93.7	77.2	2.84	10.4
√				93.8	93.4	96.0	78.8	2.96	11.4
	√			93.1	92.9	95.1	78.0	2.90	10.4
		√		94.9	94.3	96.4	79.5	3.29	15.5
			√	93.3	93.0	94.4	77.7	2.84	10.4
√	√			94.1	93.6	96.2	79.7	3.01	11.5
√		√		95.0	94.3	96.3	80.1	3.47	15.9
√	√	√		95.3	94.7	96.7	80.8	3.52	16.1
√	√	√	√	95.9	95.6	98.1	82.3	3.52	16.1

The ablation experimental results show that the individual introduction of each improved module positively contributes to model performance. Among them, the PAFPN_DDFE neck network alone achieves the highest accuracy gain, increasing mAP₅₀ by 3.0% and mAP₅₀₋₉₅ by 3.4%, respectively. The Focaler-Shape-IoU loss function improves mAP₅₀ by 1.6% without increasing parameters or computational complexity. Furthermore, the introduced C3k2_GBC module, requiring only 0.12 M parameters and a negligible computational overhead of 1.0 GFLOPs, improved mAP₅₀ and mAP₅₀₋₉₅ by 2.3% and 1.6%, respectively. On the other hand, the C2PSA_LRSA module, while maintaining the same total computational cost (10.4 GFLOPs) and adding only 0.06 M parameters, successfully increased mAP₅₀ by 1.4% and mAP₅₀₋₉₅ by 0.8%.

The results indicate that the proposed modules work collaboratively to improve overall performance, with the full GLD-YOLO model incorporating all modules delivering the best comprehensive results.

3.7. Visualization Results and Analysis

This paper selects four representative ore images from the dataset to visually demonstrate the improved algorithm’s performance in recognizing ore targets, as shown in Figure 7. The top row displays the original input images, and the remaining rows present the detection results of YOLOv8n-seg, YOLO11n-seg, YOLO13n-seg, and GLD-YOLO, respectively. The confidence score is marked beside each predicted category.

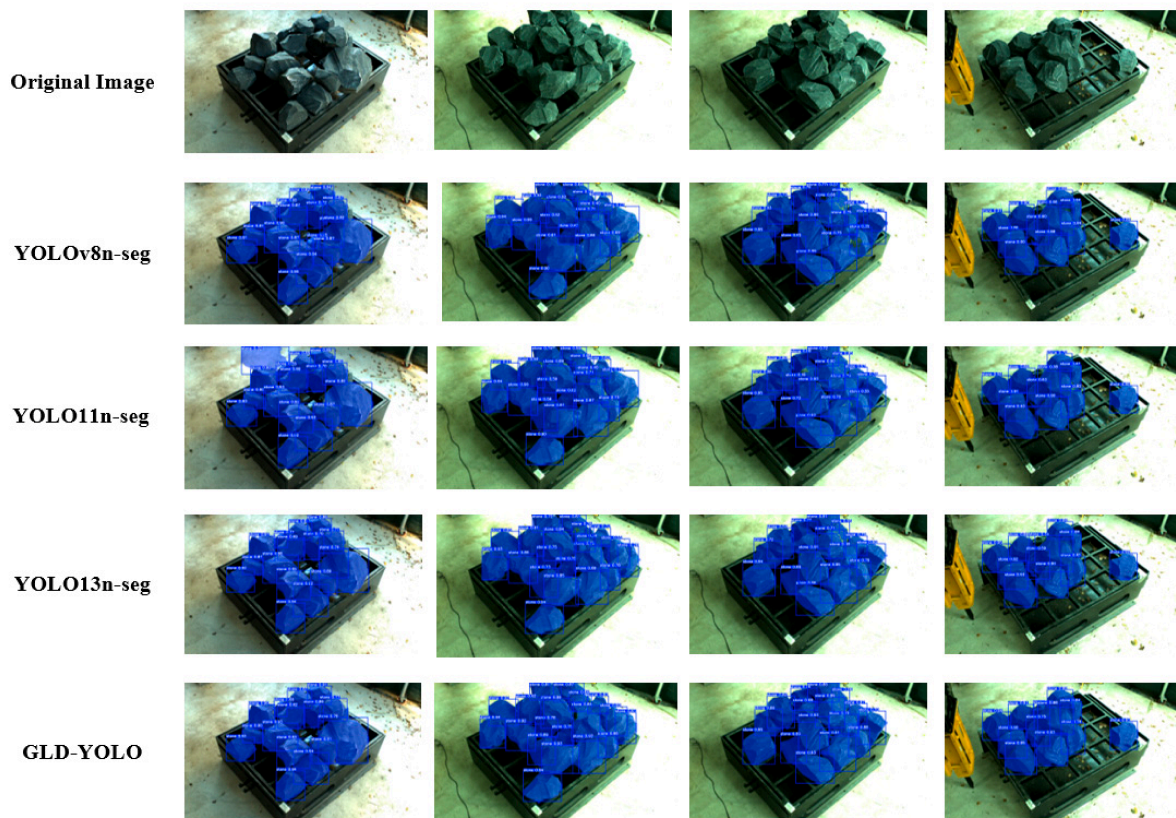


Figure 7. Comparison of detection results for different models.

As observed in Figure 7, the GLD-YOLO model achieves substantial improvements in instance segmentation accuracy and target prediction confidence in complex stacked ore scenarios. Specifically, baseline models including YOLOv8n-seg, YOLO11n-seg, and YOLO13n-seg generally suffer from over-segmentation and under-segmentation to varying degrees when handling regions with blurred edges and similar textures. For example, in the YOLO11n-seg detection results, the ground was incorrectly identified in the first image, and due to the similar texture, the two rightmost ores were incorrectly identified as the same one in the second image. Additionally, Their mask edges exhibit lower fitting accuracy and refinement compared to the proposed model.

In addition, when handling densely distributed and closely packed ores, the baseline models are prone to boundary overlap caused by inadequate edge feature extraction capability. Consequently, multiple mutually occluded individual ores are easily misidentified as a single object. Furthermore, when processing large single ores affected by non-uniform illumination or featuring multi-faceted complex textures, these baseline models often incorrectly segment one complete ore into multiple targets due to fragmented local feature representation.

3.8. Comparative Experimental Analysis of Different Improvement Strategies

3.8.1. Comparative Experiment of Different C3k2 Modules

This paper verifies the advancement and effectiveness of the C3k2_GBC module through comparative experiments, and embeds the C3k2_FasterBlock [25], C3k2_PKI [26], C3k2_ODConv [27], and C3k2_IRMB [28] structures into the YOLO11n-seg model. The results are shown in Table 4.

Table 4. Comparison experiments of different C3K2 structures.

Module	P (%)	R (%)	mAP ₅₀	mAP _{50–95}
C3k2(base)	91.1	90.9	93.7	77.2
C3k2_FasterBlock	91.6	91.1	94.2	77.5
C3k2_PKI	90.9	90.7	93.1	76.9
C3k2_ODConv	92.1	91.7	94.9	78.0
C3k2_IRMB	90.4	89.8	92.3	75.3
C3k2_GBC	93.8	93.4	95.5	78.7

Experimental results show that C3k2_GBC outperforms other modules in Precision (P), Recall (R), and mAP. This demonstrates that C3k2_GBC effectively enhances the model's fine-grained representation of local texture and geometric features, and significantly improves the network's perception of small targets and high-frequency information in ore crushing scenarios.

3.8.2. Comparative Experiment of Different C2PSA Modules

To evaluate the effectiveness of the proposed C2PSA_LRSA module in the ore instance segmentation task, several improved C2PSA structures are selected for comparative experiments. The experimental results are shown in Table 5.

Table 5. Comparison experiments of different C2PSA structures.

Module	P (%)	R (%)	mAP ₅₀	mAP _{50–95}
C2PSA(base)	91.1	90.9	93.7	77.2
C2PSA_EMA [29]	92.7	92.1	94.2	77.6
C2PSA_CBAM [30]	91.3	90.6	93.4	77.1
C2PSA_MaSA [31]	92.1	91.7	94.0	77.8
C2PSA_ECA [32]	90.7	90.1	92.3	76.8
C2PSA_LRSA	93.1	92.9	95.1	78.5

Experimental results demonstrate that C2PSA_LRSA achieves the best performance across all metrics among the compared C2PSA structures. Specifically, while C2PSA_MaSA and C2PSA_EMA deliver improvements, they involve increased parameters and computational complexity, and are inferior to C2PSA_LRSA in terms of flexibility and adaptability. In contrast, the remaining modules exhibit performance degradation to varying degrees, failing to achieve the expected segmentation gains.

3.8.3. Comparative Experiment of Different Neck Networks

To verify the impact of the proposed neck network embedded with the dual-domain feature extraction module (PAFPN_DDFE) on model performance, different feature fusion network structures are substituted into the YOLO11-seg framework for comparative experiments on the ore instance segmentation task. The experimental results are presented in Table 6.

Table 6. Comparison of different neck networks.

Module	P (%)	R (%)	mAP ₅₀	mAP _{50–95}
PAFPN(Base)	91.1	90.9	93.7	77.2
BIFPN [33]	90.9	90.4	93.4	77.0
MAFPN [34]	91.7	91.1	94.5	78.3
EMBSFPN [35]	89.8	89.6	91.1	76.7
PAFPN_DDFE	94.9	94.3	96.7	80.6

The results in Table 6 show that compared with the baseline PAFPN neck, BIFPN and EMBSFPN exhibit a slight performance degradation. Only the MAFPN network achieves a marginal improvement, with no significant variation across its Precision, Recall, and mAP metrics. This indicates that these structures have limited effectiveness for the instance segmentation of irregular ores in the specific scenario of this study. In contrast, the proposed PAFPN_DDFE demonstrates superior performance across all evaluation metrics. This confirms that the efficient shallow feature extraction and deep feature injection mechanism of PAFPN_DDFE effectively boosts detection accuracy and localization capability in dense scenarios.

3.8.4. Comparative Experiment of Loss Functions

To verify the effectiveness of the improved Focaler-Shape-IoU proposed in this study, it is compared with several other loss functions, including DIoU, GIoU, EIoU, WIoU, Shape-IoU, and Focaler-CIoU. As shown in Table 7, the results demonstrate that Focaler-Shape-IoU achieves superior performance across all evaluation metrics. These findings indicate that enhancing the model's attention to complex samples effectively improves the regression accuracy of target bounding boxes.

Table 7. Comparison experiments of loss function.

Module	P (%)	R (%)	mAP ₅₀	mAP ₅₀₋₉₅
CIoU(Base)	91.1	90.9	93.7	77.2
DIoU	90.5	90.1	92.6	76.4
GIoU	90.6	90.0	93.4	77.0
EIoU	89.5	89.1	91.2	75.2
WIoU	89.9	89.3	91.9	75.7
Shape-IoU	92.1	91.5	93.9	77.6
Focaler-CIoU	91.0	90.6	93.4	77.5
Focaler-Shape-IoU	93.6	93.3	95.3	78.2

3.9. Model Generalization Performance Evaluation

Two public datasets with distinct feature distributions, namely MinerallImage5k and Crack-seg, are adopted alongside the self-built dataset to validate the generalization ability and cross-scene adaptability of the proposed model. Rigorous comparative tests are conducted on YOLOv8n-seg, YOLO11n-seg, RF-DETR, and the proposed algorithm, aiming to evaluate the model's feature extraction and segmentation performance for targets of different scales and morphologies. The experimental results are presented in Table 8.

Table 8. Generalization experiments.

Dataset	Module	P (%)	R (%)	mAP ₅₀	mAP ₅₀₋₉₅
MinerallImage5k	YOLOv8n-seg	81.1	76.3	79.5	53.4
	YOLO11n-seg	82.6	78.0	81.1	55.4
	RF-DETR	83.8	78.9	82.3	58.1
	GLD-YOLO	84.3	79.8	83.0	58.6
Crack-seg	YOLOv8n-seg	83.2	72.1	82.5	61.4
	YOLO11n-seg	84.7	73.7	84.3	63.2
	RF-DETR	85.8	75.3	85.7	64.9
	GLD-YOLO	86.2	75.9	86.4	65.8

Experimental results show that the GLD-YOLO network proposed in this study achieves excellent cross-domain generalization performance. On the MinerallImage5k

dataset, compared with the baseline models, the Precision, Recall, and mAP_{50} of the proposed algorithm increase by 1.7%, 1.8%, and 1.9%, respectively. Furthermore, on the Crack-seg dataset characterized by complex structures and irregular edges, the model still delivers stable performance improvements, with the three aforementioned metrics rising by 1.5%, 2.2%, and 2.1%, respectively.

Furthermore, compared with the mainstream RF-DETR architecture, GLD-YOLO also demonstrates a notable accuracy advantage, achieving a 0.7% mAP_{50} improvement on both datasets. On the two datasets, it outperforms RF-DETR by 0.5% and 0.9% in terms of mAP_{50-95} , the experimental results are shown in Figure 8. In summary, GLD-YOLO demonstrates high algorithmic robustness when handling complex visual tasks, such as high-frequency texture extraction and irregular boundary segmentation. To more intuitively demonstrate the practical application effect of the model, Figure 8 shows the segmentation results of GLD-YOLO on the two datasets mentioned above. As can be seen from the figure, the model can achieve relatively accurate boundary fitting and pixel-level segmentation, whether dealing with the complex texture of mineral images or the irregular direction of cracks.

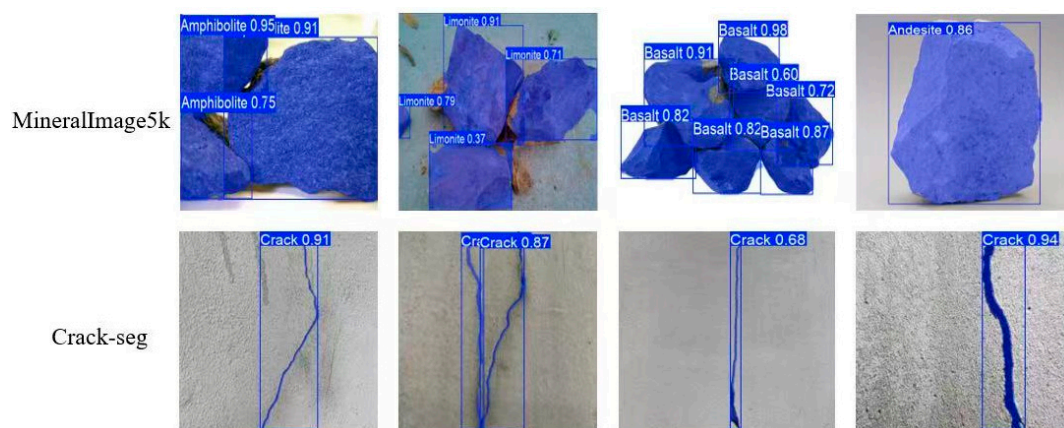


Figure 8. Segmentation Results of Generalization Experiments.

4. Conclusions

Addressing the challenges of ore instance segmentation in underground crushing scenarios—such as dense mutual occlusion, complex target shapes, and blurred boundaries—this study conducts an in-depth investigation into complex feature extraction and multi-scale feature fusion strategies. The main conclusions are as follows:

- (1) This paper proposes a multi-scale feature fusion instance segmentation algorithm, GLD-YOLO. To address the feature extraction bottleneck of conventional networks in densely stacked ore scenarios, we embed the GBC and LRSA modules to suppress environmental noise and enable fine-grained extraction of ore detail features. Furthermore, we design a PAFPN_DDFE neck network, in which the DDFE module serves as a feature filtering gateway. It leverages shallow-level features to achieve spatial-frequency dual-domain denoising and boundary enhancement simultaneously. Finally, the Focaler-Shape-IoU loss mechanism is adopted to optimize the regression of hard-to-regress bounding boxes. Compared with the baseline model, the proposed model improves mAP_{50} and mAP_{50-95} by 4.4 percentage points and 5.1 percentage points, respectively. It achieves high-precision localization and mask segmentation for irregular and heavily overlapping ores while maintaining an inference speed of 90 FPS.
- (2) The effectiveness of the proposed algorithm was verified using a self-built simulated ore crushing test platform. Results show that, compared with baseline models and

other mainstream instance segmentation algorithms, GLD-YOLO significantly improves segmentation accuracy and bounding box regression quality while maintaining good inference efficiency. Furthermore, experiments on the public datasets MinerallImage5k and Crack-seg demonstrate that the proposed model has considerable cross-scene generalization ability, achieving mAP₅₀ improvements of 1.9% and 2.1% respectively over the baseline, thereby validating the rationality and effectiveness of the model improvements.

- (3) This study explores algorithms for intelligent crushing scenarios. The proposed model not only achieved ideal theoretical analysis results and experimental performance, but the related network structure optimization ideas also provide a valuable reference for algorithm iteration in the field of ore image segmentation. Despite the overall positive results, several limitations remain: First, the current model's performance evaluation is based solely on a simulation platform and has not yet been field-tested in a real mine environment; second, while the structural optimization scheme introduced in this paper enhances the model's capabilities, it also increases the number of model parameters to some extent. To address these shortcomings, future research will focus on three directions: developing lightweight models to reduce computational complexity, collecting real-world datasets from mine working faces for field verification, and achieving deep integration of visual segmentation algorithms with the crushing robot control system, ultimately constructing a perception-execution closed-loop control framework.

Author Contributions: Conceptualization, M.C. and L.S.; methodology, Y.L.; software, Y.L., X.W. and J.G.; validation, Y.L., L.S., X.W. and J.G.; formal analysis, M.C., L.S. and X.W.; investigation, Y.L. and J.G.; resources, M.C. and X.W.; data curation, Y.L. and L.S.; writing—original draft preparation, Y.L.; writing—review and editing, M.C. and L.S.; visualization, Y.L.; project administration, X.W.; funding acquisition, M.C. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by State Key Laboratory Special Programs of China Minmetals Corporation (No. 2024GZKJ02).

Data Availability Statement: Data are contained within the article; the code is confidential.

Conflicts of Interest: Authors Miao Chen and Xitao Wang were employed by Changsha Institute of Mining Research Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Lampinen, S.; Koivumäki, J.; Mattila, J. Full-Dynamics-Based Bilateral Teleoperation of Hydraulic Robotic Manipulators. In *Proceedings of the 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), Munich, Germany, 20–24 August 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 1343–1350. [\[CrossRef\]](#)
2. Wang, G. Latest technical progress and discussion on problems of coal mine intelligentization. *Coal Sci. Technol.* **2022**, *50*, 1–27. [\[CrossRef\]](#)
3. Chen, C.; Li, H.; Zhu, W.; Niu, L. Blast pile fragments recognition algorithm for open pit mines based on deep learning and nap-of-the-object photogrammetry. *J. Min. Sci. Technol.* **2025**, *10*, 151–162. [\[CrossRef\]](#)
4. Lampinen, S.; Niu, L.; Hulttinen, L.; Niemi, J.; Mattila, J. Autonomous robotic rock breaking using a real-time 3D visual perception system. *J. Field Robot.* **2021**, *38*, 980–1006. [\[CrossRef\]](#)
5. Zhang, F.; Zhang, J.; Cheng, H. Research review on intelligent object detection technology for coal mines based on deep learning. *Coal Sci. Technol.* **2025**, *53*, 284–296.
6. Yang, L.; Cao, C.; Liu, S. Research on Ore Segmentation and Location Based on Improved Mask RCNN Algorithm. *Min. Res. Dev.* **2023**, *43*, 209–216. [\[CrossRef\]](#)

7. Luo, X.; Liu, S.; Tang, W.; Wang, X. Research on identification and location of blocked ore at silo inlet based on Mask RCNN. *Nonferrous Met. Sci. Eng.* **2022**, *13*, 101–107. [\[CrossRef\]](#)
8. Shrivastava, S.; Bhattacharjee, S.; Deb, D. Segmentation of mine overburden dump particles from images using Mask R CNN. *Sci. Rep.* **2023**, *13*, 2046. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. [\[CrossRef\]](#)
10. Jin, H.; Weng, C. Foreign object detection algorithm for coal mine conveyor belt based on LDAFL-YOLOv11. *Coal Sci. Technol.* **2026**, 1–14. Available online: <https://link.cnki.net/urlid/11.2402.td.20251201.1504.004> (accessed on 10 March 2026).
11. Qiu, Z.; Huang, X.; Sun, Z.; Li, S.; Wang, J. GS-YOLO-Seg: A Lightweight Instance Segmentation Method for Low-Grade Graphite Ore Sorting Based on Improved YOLO11-Seg. *Sustainability* **2025**, *17*, 5663. [\[CrossRef\]](#)
12. Lv, Z.; Sha, T.; Sun, M.; Tu, Y.; Liu, Y.; Shao, M.; Wang, X.; Yu, L.; Wang, W. Research on foreign object detection and sorting strategy in raw coal based on an improved YOLACT. *Int. J. Coal Prep. Util.* **2025**, 1–29. [\[CrossRef\]](#)
13. Zhang, K.; Wang, Q.; Zou, G.; Yang, J.; Song, T.; Liu, Y.; Liu, D. An efficient anchor-free model for ore particle size detection. *Eng. Appl. Artif. Intell.* **2025**, *162*, 112304. [\[CrossRef\]](#)
14. Zhang, J.; Guo, X.; Liu, Y.; Shi, W.; Liu, Y.; Guan, S. Coal gangue identification method based on SDSE-YOLO in complex operating conditions. *Coal Sci. Technol.* **2026**, 1–20.
15. Jocher, G.; Qiu, I. Ultralytics YOLO11. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 10 March 2026).
16. Hua, B.-S.; Tran, M.-K.; Yeung, S.-K. Pointwise convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 19–21 June 2018. [\[CrossRef\]](#)
17. Tian, Z.; Wu, J.; Zhang, W.; Chen, W.; Zhou, T.; Yang, W.; Wang, S. Brightness improvement and detail enhancement method of mine low-illumination image based on Transformer and adaptive feature fusion. *Coal Sci. Technol.* **2024**, *52*, 297–310. [\[CrossRef\]](#)
18. Zhao, X.; Li, J. A coal gangue detection method based on deep learning. *J. Min. Sci. Technol.* **2021**, *6*, 730–736. [\[CrossRef\]](#)
19. Lau, K.W.; Po, L.-M.; Rehman, Y.A.U. Large separable kernel attention: Rethinking the large kernel attention design in cnn. *Expert Syst. Appl.* **2024**, *236*, 121352. [\[CrossRef\]](#)
20. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017. [\[CrossRef\]](#)
21. Wang, W.; Zhang, K.; Lü, Z.; Gu, Z.; Qian, H.; Zhang, Q. Machine vision detection of foreign objects in coal based on deep learning. *J. Min. Sci. Technol.* **2021**, *6*, 115–123. [\[CrossRef\]](#)
22. Nesteruk, S.; Agafonova, J.; Pavlov, I.; Gerasimov, M.; Latyshev, N.; Dimitrov, D.; Kuznetso, A.; Kadurin, A.; Plechov, P. MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description. *Comput. Geosci.* **2023**, *178*, 105414. [\[CrossRef\]](#)
23. Liu, B.; Kang, J.; Guan, H.; Ma, L.; Yu, Y.; Li, S.; Li, J. Boundary-Guided Real-Time Semantic Segmentation and Pixel-Level Quantification of Pavement Cracks. *IEEE Trans. Intell. Transp. Syst.* **2025**, *27*, 3459–3475. [\[CrossRef\]](#)
24. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
25. Dong, J.; Deng, L.; Wan, D.; Liu, C.; Yin, J.; Guo, M.; Zhang, H.; Lin, S.; Liu, H.; Liu, L. Oriented bounding box detection algorithm for dense scenarios of robotic arm operation. *Expert Syst. Appl.* **2025**, *298*, 129678. [\[CrossRef\]](#)
26. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly kernel inception network for remote sensing detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024. [\[CrossRef\]](#)
27. Li, C.; Zhou, A.; Yao, A. Omni-dimensional dynamic convolution. *arXiv* **2022**, arXiv:2209.07947. [\[CrossRef\]](#)
28. Zhang, J.; Li, X.; Li, J.; Liu, L.; Xue, Z.; Zhang, B.; Jiang, Z.; Huang, T.; Wang, Y.; Wang, C. Rethinking mobile block for efficient attention-based models. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; IEEE: Piscataway, NJ, USA, 2023. [\[CrossRef\]](#)
29. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023. [\[CrossRef\]](#)
30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018. [\[CrossRef\]](#)
31. Fan, Q.; Huang, H.; Chen, M.; Liu, H.; He, R. Rmt: Retentive networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024. [\[CrossRef\]](#)
32. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020. [\[CrossRef\]](#)
33. Doherty, J.; Gardiner, B.; Kerr, E.; Siddique, N. BiFPN-YOLO: One-stage object detection integrating Bi-Directional feature pyramid networks. *Pattern Recognit.* **2025**, *160*, 111209. [\[CrossRef\]](#)

34. Chang, H.; Bassir, D.; Barrios, A.; Chen, G. Fourier-Mixture of Experts YOLO for concrete crack segmentation with visual interpretability. *Autom. Constr.* **2025**, *179*, 106452. [[CrossRef](#)]
35. Yang, Z.; Guan, Q.; Yu, Z.; Xu, X.; Long, H.; Lian, S.; Hu, H.; Tang, Y. Mhaf-yolo: Multi-branch heterogeneous auxiliary fusion yolo for accurate object detection. *arXiv* **2025**, arXiv:2502.04656. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.