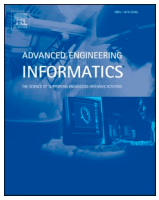
Advanced Engineering Informatics 71 (2026) 104319

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Full length article

From laboratory to field: normal map-aided multimodal instance segmentation for blasting fragmentation analysis

Yulin Wang [a,d], Xin Wang [d], Yudi Tang [a,*], Xu Dai [c], Jinming Dong [e], Guangyao Si [a,b,*]

[a] School of Minerals and Energy Resources Engineering, The University of New South Wales, Sydney, NSW 2052, Australia
[b] State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, Chengdu 610059, China
[c] China Construction Eighth Engineering Division Corp., LTD, 200112 Shanghai, China
[d] School of Civil Engineering, Southeast University, Nanjing 210096, China
[e] AIM Mining Corporation, Perth, WA 6005, Australia

ARTICLE INFO

Keywords:
Normal map
Multimodal deep learning
Instance Segmentation
Blasting Fragmentation
Size distribution analysis

ABSTRACT

Particle size analysis of rock fragments plays a crucial role in mining engineering. However, traditional non-contact image-based methods typically rely solely on RGB images, which are highly sensitive to illumination changes, shadow interference, and fragment texture. This reliance limits the accuracy and generalization capability of conventional approaches and often necessitates retraining when applied across different scenes. To address these issues, this study introduces normal maps and provides a detailed analysis of their advantages in representing rock fragment features. Furthermore, we propose a multimodal instance segmentation framework named Adaptive Feature Recombination Network (AFRNet). AFRNet incorporates a modality effectiveness perception mechanism to adaptively guide the fusion process while suppressing interference from unreliable modalities. In addition, it employs a multi-scale attention fusion module to fully exploit and utilize the strength of each modality. This study systematically compares three fusion strategies—data-level, feature-level, and decision-level—and conducts experiments under various modality combinations. Experimental results demonstrate that incorporating normal maps significantly improves segmentation accuracy and enhances model robustness in degraded environments such as low illumination and shadow interference. Moreover, the model trained in a laboratory environment is directly transferred, without retraining, to a practical particle size analysis task at an actual mining site in Nanjing. The resulting particle size distribution curves exhibit a deviation of less than 10% compared with manually labeled results, validating the proposed method's zero-cost transferability and engineering applicability.

1. Introduction

Rock fragmentation is a key process in mining and civil engineering activities, involving techniques such as drilling and blasting to break rock masses from their in-situ condition or large blocks into smaller fragments [1]. Rock fragmentation affects various aspects of mining and excavation operations, including excavation productivity, material handling efficiency, downstream processing, and the quality of extracted ore. To achieve optimal blasting results, it is crucial to study the distribution patterns of fragmented rock sizes. Traditional evaluation of rock fragment sizes relies on manual sieving or visual estimation, both of which are time-consuming and labor-intensive. As a result, automated analysis of fragment sizes has become a major focus for many researchers and scholars [2–4]. At present, the vast majority of automatic rock fragment extraction methods rely on single-channel RGB images captured on-site and employ machine learning or deep learning models. Single-modality segmentation methods are easily affected by the texture, color, and illumination conditions of rock fragments, and often require retraining when transferred across engineering scenarios, thereby limiting both segmentation performance and cross-scenario generalization.

To address the challenges of poor cross-scene transferability, insufficient utilization of multimodal information, this study introduces the normal map modality, highlights its advantages in representing rock features, and develops a multimodal feature-level fusion instance segmentation model specifically designed to leverage this modality. This

* Corresponding authors at: School of Minerals and Energy Resources Engineering, The University of New South Wales, Sydney, NSW 2052, Australia.
E-mail addresses: wangyl025@163.com (Y. Wang), xin.wang@seu.edu.cn (X. Wang), yudi.tang@unsw.edu.au (Y. Tang), daixu198500@foxmail.com (X. Dai), j.dong@aimmining.com.au (J. Dong), g.si@unsw.edu.au (G. Si).

https://doi.org/10.1016/j.aei.2026.104319
Received 2 August 2025; Received in revised form 22 November 2025; Accepted 5 January 2026

study constructed a multimodal dataset comprising RGB images, depth maps, and normal maps (RGB-D-N) using a fragment sampling box and photogrammetry techniques. The normal map was introduced as a novel geometric modality. Based on this dataset, a novel multimodal fusion instance segmentation model—Adapt Feature Recombination Net (AFRNet)—was proposed. Various fusion strategies—including data-level, feature-level, and decision-level fusion—were systematically compared in terms of segmentation performance, along with an analysis of the advantages and disadvantages of different modality combinations. The results were benchmarked against several state-of-the-art instance segmentation models, including Mask2Former, SOLOv2, YOLACT, IAM and SAM [5–9], to demonstrate the superiority of AFRNet. AFRNet, which was trained on laboratory data, was applied to a mining site in Nanjing, where multimodal data were collected for prediction. Particle size distribution and granulometric analysis were then conducted, and the results were compared with those obtained using WipFrag software and manual annotations. The comparison further validated the cross-scenario transferability and engineering applicability of AFRNet with the inclusion of the normal map modality.

1.1. Literature review

In recent years, machine learning and deep learning methods have been increasingly applied in the field of civil engineering [10–15]. In civil engineering, image-based particle analysis methods are gaining widespread use. Classical image processing techniques such as edge detection, and software developed from these methods—like Wipfrag and Split-Online—can automatically extract rock fragment information [16,17]. However, these methods often lack robustness and are only effective when fragment boundaries are clear and image ambiguity is low. Current mainstream research in rock fragmentation primarily focuses on semantic segmentation methods [18–20]. Zhong et al. [21] compared classical semantic segmentation models such as UNet, SegNet, and PSPNet, and evaluated their performance through morphological analysis. Nevertheless, due to occlusion and shadowing in rock images, adjacent rock fragments often do not share the same pixel boundaries, making it difficult to distinguish their edges. This often necessitates the integration of additional image processing methods. Qiao et al. [22] applied instance segmentation to rock fragment identification in tunnels, demonstrating that instance segmentation is more effective than semantic segmentation. Dai et al. [23] conducted a comparative study on the performance of classical instance segmentation networks, including Mask R-CNN, YOLACT, and Mask2Former, in the context of rock particle segmentation. They further compared the accuracy of particle count and gradation statistics derived from each model, providing valuable insights for the application of instance segmentation methods in rock fragmentation analysis. However, instance segmentation methods based solely on RGB images are susceptible to uneven lighting, shadows, and color variations. Their performance degrades significantly in scenes with high color contrast or extreme lighting conditions, revealing a need for improved robustness.

In recent years, with the advancement of 3D laser scanning and close-range photogrammetry technologies, acquiring surface geometric information of objects has become increasingly accessible and cost-effective. As an important data form for describing object geometry, depth images have been successfully applied in various multimodal computer vision tasks. In scenes with low contrast and complex backgrounds, the structural and positional features provided by depth images serve as a valuable geometric supplement to RGB images [24–27]. In face recognition tasks, Jiang et al. [28] designed modality-specific networks to learn complementary and shared features across RGB and depth modalities, transforming them into a unified feature space. Their multimodal matcher achieved strong performance in large-scale databases. The introduction of multimodal data offers additional spatial and contextual information, facilitating more robust and accurate scene understanding. Fu et al. [29] proposed a Mask R-CNN-based instance segmentation model tailored for RGB-D four-channel input for pose estimation, and conducted comparative experiments with various classical segmentation networks. The results demonstrated the significant superiority of four-channel input over conventional RGB input, highlighting its advantages in accuracy and robustness. Compared with single-modality inputs, multimodal fusion typically improves model performance, though it also introduces uncertainty due to redundant information [30,31]. RGB-D images exhibit strong anti-interference capabilities: depth channels clearly differentiate foreground from background and provide global spatial information and coarse pose estimation, offering a defined visual scope for detail-level segmentation. On the other hand, RGB images exhibit gradient changes at object edges, which help capture finer texture details and richer local features.

Current research on RGB-D image segmentation mainly focuses on how to utilize the extra depth channel to improve recognition accuracy. For example, Gupta et al. [32] and Wang et al. [33] employed two separate convolutional neural networks (CNNs) to extract RGB and depth features and performed segmentation on candidate regions of indoor objects. Similarly, Lin et al. [34] extracted RGB features via CNNs and used hierarchical discretized depth images to represent different scene resolutions, aligning image regions across network branches that share convolutional feature maps, and combining the results for final segmentation. Wang et al. [35] proposed depth-aware convolution and depth-aware average pooling methods, which seamlessly integrate geometric information into CNNs by leveraging depth similarity between pixels during information propagation. Essentially, these methods embody a dual-modality design that leverages both color and depth information. The dual-modality approach offers a richer set of raw features and, under the premise of effectively suppressing modality interference and achieving cross-modal information coupling, holds greater potential for image segmentation compared to single-modality methods [36–38].

In the field of civil engineering, Lu et al. [39] proposed a multi-scale RGB-D image fusion algorithm for segmenting multi-scale rock images on conveyor belts, and further demonstrated the superiority of this method over single-sensor image segmentation approaches. Li et al. [40] developed a segmentation network for post-blasting color images by integrating a Deformable Convolutional Network (DCN) with a Transformer Attention Mechanism (TAM). This network adaptively preserves both global and local features of the image, thereby enhancing the segmentation and recognition of rock fragment edges. Liu et al. [41] improved the HRNetv2 semantic segmentation model to segment drone-captured images of rock fragments. They then employed the Discrete Element Method (DEM) to simulate and derive upper and lower bound area gradation curves from the designed gradation envelope, enabling gradation evaluation. Fan et al. [42] proposed a model named Rock-net for rock fragment segmentation, introducing a regression model to construct a nonlinear relationship between rockpile morphology parameters and spatial gradation. Zhang et al. [43] proposed an enhanced Mask R-CNN model for segmenting rock images, and used GCNet for gradation calibration to obtain more accurate results. The method was validated on a hyperspectral database containing various lithological classes, demonstrating good effectiveness and accuracy. However, some of the aforementioned methods still rely solely on semantic segmentation and have not fully explored instance segmentation techniques. Others are limited to color images or single-modality data, lacking the integration of multimodal information. Therefore, a framework that combines instance segmentation algorithms with multimodal data may be more suitable for rock fragmentation analysis tasks. In addition, color image-based methods often face challenges in cross-scene generalization, while depth images tend to exhibit blurred features around particle edges, presenting inherent limitations. Consequently, it is necessary to introduce a novel modality to complement geometric information and enhance the accuracy of edge segmentation.

In animation modeling, the normal map is a commonly used 2D texture map that defines how lighting interacts with the material

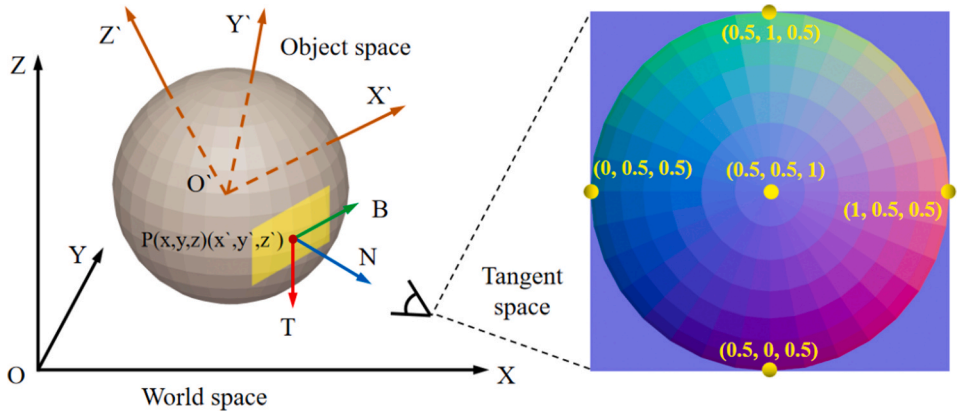


Fig.1. Illustration of the Normal Map Principle and Numerical Example.

represented by each pixel [44]. A normal map stores the X, Y, and Z components of surface normals as pixel values in a regular 2D image, providing surface geometric information distinct from depth images. Since normal map pixel values are independent of the corresponding region's position within the global model, they are advantageous for representing local details compared to depth images. Tang et al. [45] applied semantic segmentation to extract rock fragments by concatenating RGB images with normal maps along the channel dimension, and compared the results with traditional single-modality methods to analyze the advantages. However, this study did not fully explore instance segmentation methods, and the use of simple channel concatenation to fuse different modalities fails to effectively leverage the complementary information between modalities, presenting certain limitations.

1.2. Contributions of our work

The key contributions can be summarized as the following three points:

(i) **Introducing normal maps into the rock fragment recognition task and analyzing their advantages.** This study systematically compares the pixel-value representations of RGB images, normal maps, and depth maps in rock fragment imagery. Through geometric principles and pixel matrix distribution analysis, it demonstrates the significant advantages of normal maps in feature representation.

(ii) **Proposing AFRNet, an instance segmentation model designed for RGB–N multimodal input.** AFRNet introduces modality feature weight parameters, an Adaptive Reassembly Module, and a Scale-Progressive Multi-Modal Fusion module to achieve multimodal feature interaction. Performance comparisons with classical algorithms further validate the effectiveness of AFRNet.

(iii) **Enhancing the generalization capability of image-based rock fragment recognition methods.** Benefiting from the stability of normal map features, the proposed method outperforms conventional approaches in degraded scenarios such as low illumination and shadow interference. Moreover, model weights trained in laboratory environments can be directly applied to mine-site scenarios without retraining, yielding reliable fragment analysis results.

2. Acquisition of Multi-Modal image data

2.1. Definition of normal maps

In the fields of animation modeling and computer graphics, normal maps are textures used to store the surface normal directions of 3D models [46]. They are commonly employed to compute the reflection angles of light on a model's surface and to simulate bumps and lighting effects. This allows a 3D model to exhibit more detailed lighting and shading without increasing the number of polygonal faces. Normal maps are typically stored as standard RGB images, where the R, G, and B components correspond to the X, Y, and Z components of the surface normal, respectively. Since the normal vector components range from [-1, 1], while RGB values range from [0, 1], a transformation is required to map each normal component into the RGB range:

$$C_{r,g,b} = 0.5 \times N_{x,y,z} + 0.5 \tag{1}$$

Here, $C_{r,g,b}$ represents the RGB values, and $N_{x,y,z}$ denotes the normal vector components. Similarly, when reading normal data from a normal map, the inverse transformation must be performed to restore the original normal values.

The pixel values in a normal map differ depending on the coordinate system in which the normals are defined, and there exist specific spatial transformations among these coordinate systems. As shown in Fig. 1, the world space refers to the global coordinate system where light sources and the observer are located; the object space refers to the local coordinate system of the object itself, containing information related to the model's vertices; and the tangent space refers to the independent space of each model surface's texture, also known as the texture space.

In object space, normal information is computed relative to the object itself, resulting in normal maps with rich colors due to the variety of direction vectors. In tangent space, normals are computed relative to the surface tangent and generally point close to the direction (0,0,1) (0, 0, 1) (0,0,1), which corresponds to an RGB value of approximately (0.5,0.5,1) (0.5, 0.5, 1) (0.5,0.5,1). Therefore, normal maps in tangent space typically appear bluish in overall tone. The RGB values of different pixel positions are labeled accordingly in the figure.

For object-space normal maps, pixel values are defined relative to a specific model. When applied to different objects, discrepancies in object transformation matrices can lead to distorted normal information. In contrast, tangent space normal maps encode normal directions relative to the local surface, thus exhibiting better generalizability and adaptability to different models. In our previous work, we compared the performance of object-space and tangent space normal maps on rock fragment models, demonstrating the advantages of tangent space normal maps [45]. Therefore, all normal maps employed in this study are defined in tangent space.

2.2. Advantages of normal maps in representing rock fragment features

In our analysis of two real-world mine site images, we observed a common yet often overlooked phenomenon: the RGB values of many pixels are remarkably similar across all three channels, resulting in an

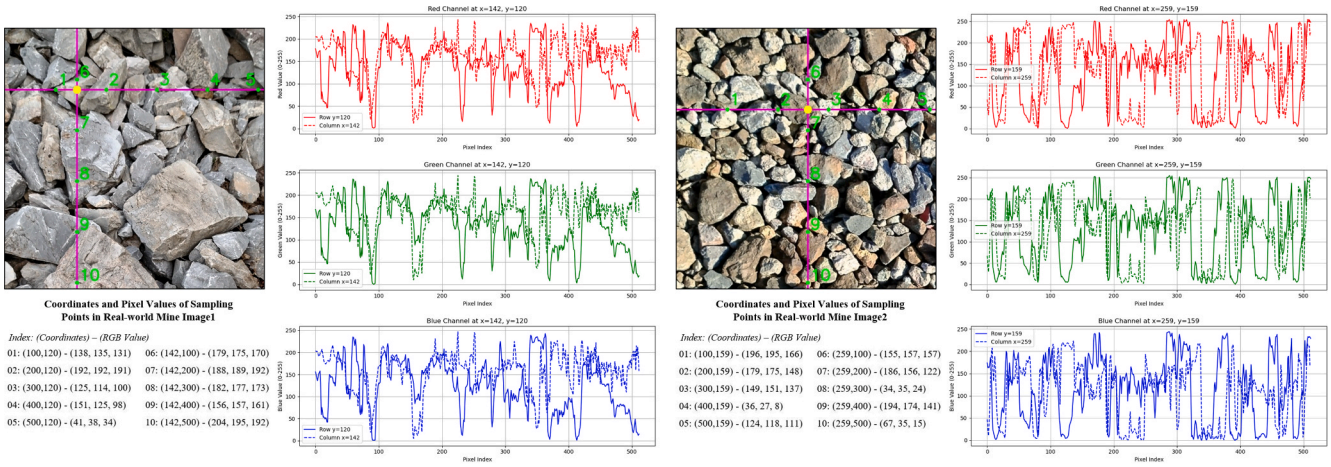


Fig.2. RGB Distribution and Intensity Profiles of Sampling Points in Real-world Mine Images.

overall distribution that closely resembles a grayscale image. To further validate the generality and consistency of this phenomenon, as shown in Fig. 2, we selected representative regions and extracted pixel values along both horizontal and vertical directions. The visualization of channel-wise intensity profiles revealed that the RGB values not only share similar magnitudes but also exhibit highly consistent trends across different spatial positions, indicating a lack of color diversity and low inter-channel variance in the image. We hypothesize that this phenomenon arises from a combination of factors:

First, from the perspective of physical imaging mechanisms, the surfaces of fragmented rocks in open-pit mines are typically composed of granular materials with relatively homogeneous mineral compositions and varying particle sizes. These surfaces exhibit minimal reflectance variation within the visible spectrum, lacking distinct spectral features. As a result, the incident light generates nearly identical responses in the R, G, and B channels, leading to very similar intensity values recorded by the image sensor.

Second, under natural illumination conditions or constrained field environments, images are often captured using diffuse daylight or automatic exposure in low-light scenarios. This reduces the color dynamic range and saturation. Furthermore, the camera's Auto White Balance (AWB) mechanism applies linear or nonlinear normalization across RGB channels, further compressing inter-channel color differences and driving the image toward a neutral gray appearance.

Third, from a signal processing pipeline standpoint, some images may have undergone lossy compression (e.g., JPEG encoding), during which color space conversion to YCbCr and chroma subsampling significantly diminish the original RGB distinctions. Additionally, many cameras apply default enhancement algorithms—such as color equalization, denoising, or gamma correction—which, while improving visual quality, may inadvertently smooth out subtle color variations across channels.

In summary, the near-identical RGB values observed in fragmented rock images result from a combination of physical surface properties, environmental lighting conditions, and color-preserving limitations in image processing pipelines. This characteristic deserves particular attention in subsequent tasks such as multimodal fusion or image enhancement, as it may impair a deep learning model's ability to capture color-related discriminative features, thus affecting overall representational effectiveness.

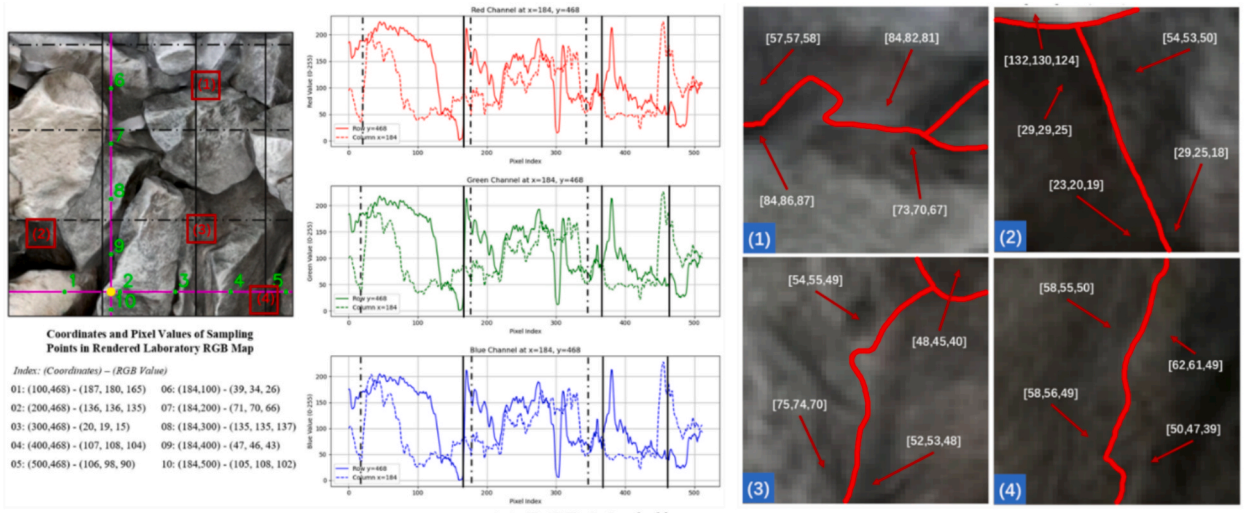
We collected fragmented rock samples of various sizes and colors, and conducted 3D reconstruction under different lighting conditions in a controlled laboratory environment to generate corresponding color texture maps. Experimental results show that these color maps consistently exhibit highly similar values across the R, G, and B channels, resulting in an overall grayscale-like appearance. This indicates that the issue is not limited to real-world mine site images, but also persists in high-precision laboratory-generated models.

In contrast, normal maps do not exhibit such inter-channel similarity. The R, G, and B channels of a normal map respectively encode the X, Y, and Z components of surface normals in 3D space. These values are derived from geometric directional information rather than surface color or reflectance intensity. As a result, the three channels of a normal map inherently maintain statistical independence and directional specificity. This geometric constraint enables normal maps to provide more distinctive and structurally informative features in tasks such as image enhancement and multi-channel feature extraction, particularly enhancing performance in boundary delineation and spatial morphology reconstruction of fragmented rocks.

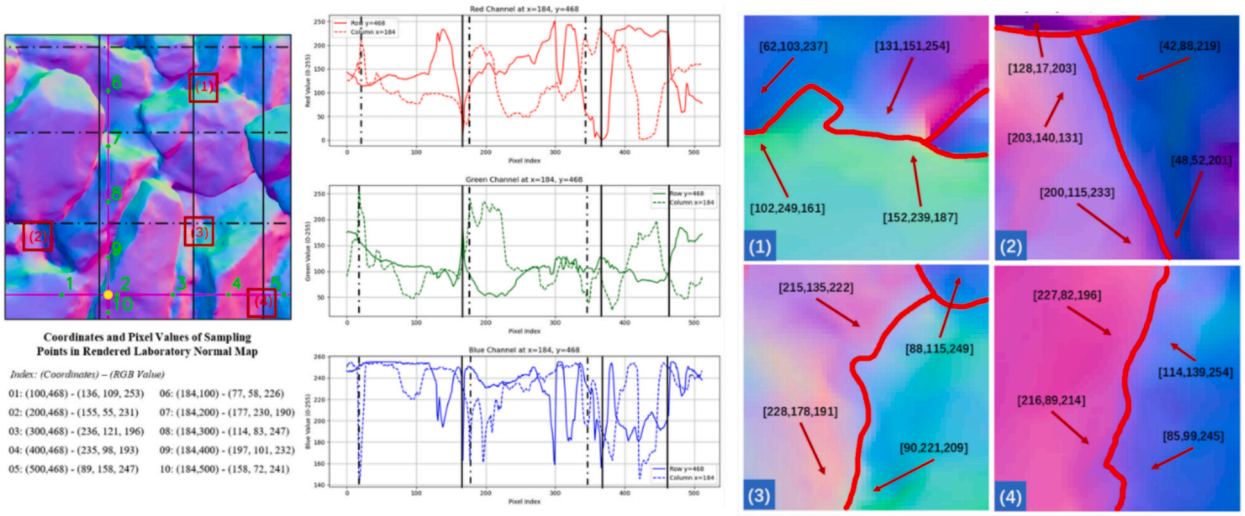
Fig. 3 presents the RGB or grayscale values of sampling points from three different modalities—color images, normal maps, and depth maps—along with their pixel intensity trends in both horizontal and vertical directions. sampling points are marked with green dots, and their coordinates and pixel values are annotated directly on the images. on the right side of the figure, solid lines represent pixel sequences along the horizontal direction (image rows), while dashed lines correspond to vertical direction (image columns). to strengthen the correspondence between image content and line plots, black solid and dotted lines are added to indicate the spatial linkage between image coordinates and plotted curves.

From this visualization, significant differences can be observed among the modalities in representing particle boundaries and structural details. In the color images, the RGB channels exhibit highly consistent distributions, demonstrating strong inter-channel correlation, which verifies the previously mentioned "near-grayscale" phenomenon. Although intensity peaks can be observed in highlights and shadowed regions, these fluctuations are primarily caused by lighting, reflections, and environmental shadows, rather than being aligned with actual particle edges. Furthermore, in regions where lighting transitions are smooth, the color image fails to adequately depict structural boundaries.

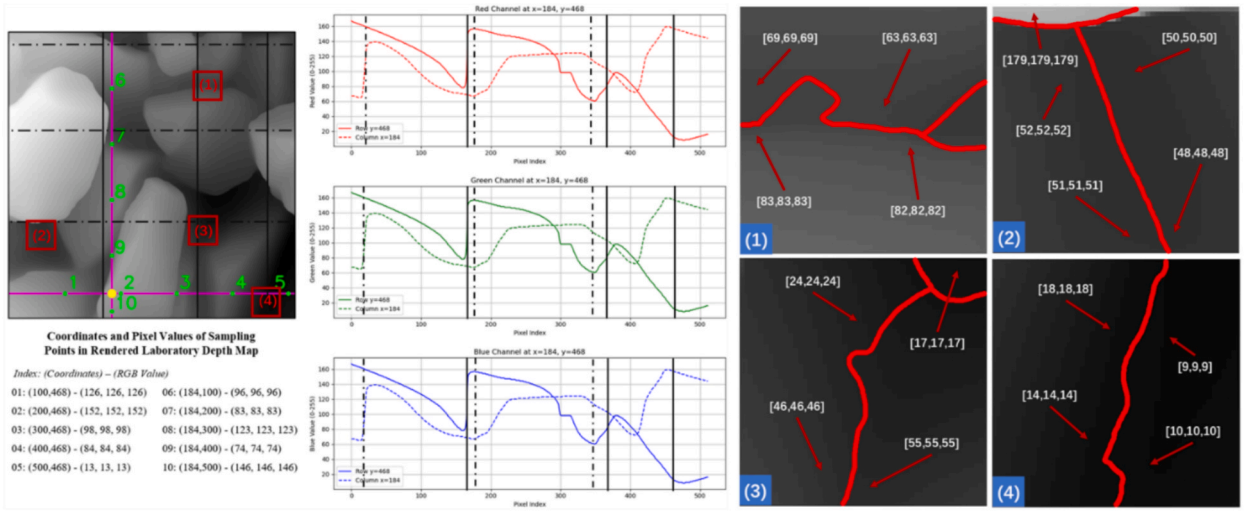
In contrast, normal maps encode the surface normal components along the X, Y, and Z directions into the R, G, and B channels, respectively, leading to inherent directional differences among the channels. Rather than exhibiting uniform intensity, each channel reflects local geometric variations in different spatial directions. At particle edges, certain channels often display sharp changes or piecewise gradients that align well with physical boundaries, thus improving edge localization accuracy and robustness. Within relatively flat interior regions of particles, the channels show more consistent and smooth value distributions, which help preserve instance connectivity and support the



(a) RGB Modality



(b) Normal Modality



(c) Depth Modality

Fig.3. Visualization of Pixel Distributions and Response Characteristics of Sampling Points in Three Modalities, Green circles indicate the locations of the sampling points, with their coordinates and corresponding pixel values annotated on the images. In the middle line charts, solid lines represent pixel sequences along the horizontal direction (image rows), and dashed lines represent variations along the vertical direction (image columns). Black solid and dashed lines indicate the corresponding positions of the sampling points in the images and the plots. Panels (1), (2), (3), and (4) show zoomed-in views of the sampled local regions in different modalities.

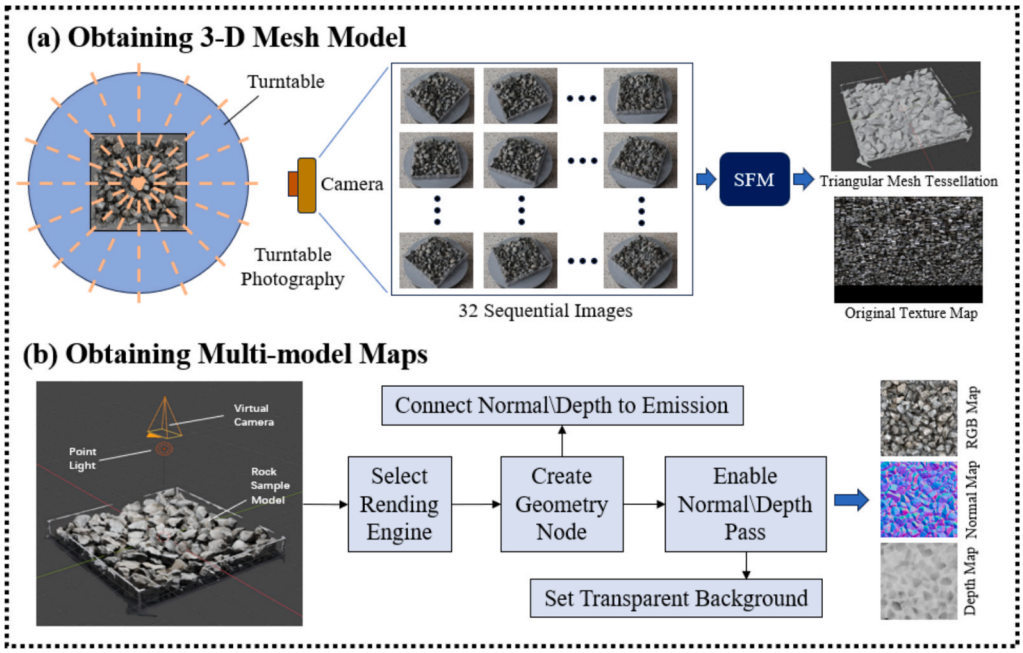


Fig. 4. Workflow for Multi-Modality Dataset Acquisition.

formation of complete, enclosed object representations in segmentation tasks.

However, we also observe that in edge regions with significant geometric variation, the three channels in the normal map do not always respond uniformly—some show strong fluctuations while others remain stable. This channel-wise inconsistency suggests that certain channels may be less informative or even introduce noise in specific regions. Therefore, in the design of subsequent multimodal feature fusion modules, it is necessary to incorporate not only spatial attention mechanisms but also channel attention strategies, enabling the model to adaptively assign weights to each channel based on local context. This approach helps emphasize informative channels while suppressing misleading or redundant signals. This finding underscores the interpretability of normal maps as a structure-aware modality and provides valuable insights for optimizing network architectures.

The depth map, although offering a concise representation of surface geometry, typically encodes pixel values proportional to the distance from the camera to the object surface. This allows it to avoid color and lighting interference. However, due to the presence of high-frequency surface undulations and occlusions in rock fragment scenes, depth maps often suffer from missing or blurred information at particle boundaries, limiting their effectiveness in fine-grained boundary detection and instance-level segmentation.

In addition, we sampled and zoomed in on four particle edge regions to illustrate pixel variations across different modalities. It can be observed that, due to the relatively uniform color tones in the color images, surface textures often interfere with the accurate delineation of particle edges, particularly for dark-colored particles. Depth maps perform well at edges of particles with suspended surfaces; however, for most particles that are relatively flat and have little height variation, the depth representation of their edges is poor. In contrast, normal maps effectively capture the angular changes at particle junctions and are not affected by surface texture or particle size variation. The four sampled regions shown in Fig. 3 clearly demonstrate how normal maps depict pixel variations at particle edges.

In summary, the analysis in Fig. 3 demonstrates that normal maps outperform color and depth images in representing both the structural edges and overall shape of rock particles. As such, normal maps offer stronger discriminative power and structural awareness in multimodal segmentation tasks, making them a valuable modality for feature extraction and fusion.

2.3. Data acquisition process

The workflow for acquiring multimodal image data is illustrated in Fig. 4, which consists of two main steps:

(1) Multi-angle inclined images of the sample box are captured using a fixed camera and a turntable. Then, stereo vision algorithms are applied to the multi-view image sequences to reconstruct a 3D point cloud of the fragment surface. Finally, a digital surface model (DSM) is obtained through triangulated mesh generation.

(2) By cropping the model, configuring global illumination, setting orthographic camera pose and viewing angles, rendering is performed to generate the diffuse map (containing surface color information), the normal map (encoding geometric details), and the depth map (representing elevation or distance information).

To ensure sufficient diversity in the dataset, the rock fragment samples were designed and categorized from three perspectives: color, size, and illumination. In terms of color, the pixel intensities of each rock fragment region were first normalized by linearly stretching the grayscale values to the 0–255 range, eliminating the influence of overall dark or bright conditions on color assessment. The normalized pixel colors were then analyzed using a clustering method (K-means). Based on the number and distribution of cluster centers, rock fragments were classified as monochromatic, bicolored, or multicolored, where the number of cluster centers directly reflects the color complexity of the fragment: more centers indicate richer colors, whereas fewer centers indicate simpler colors. Monochromatic fragments were further subdivided into light gray, dark gray, and white categories according to their average grayscale values (grayscale thresholds: 85, 170) to quantify visual differences in color.

Regarding size, fragments were divided into large, medium, and small particles based on their average area in the images (area thresholds: 121, 390 pixels), ensuring that each particle category was sufficiently represented in the dataset. For illumination, to avoid interference of fragment color on contrast statistics, the color images were converted to the HSV color space, and the V channel (brightness) of the entire image was used for statistical analysis. Fragments were then

Table 1
Dataset Diversity Partition.

Dimension	Category	Determination Method	Threshold	Sample Quantity
Color	Light gray	Cluster number + mean gray value	Cluster number = 1, mean gray value 84–170	1136
	Dark gray	Cluster number + mean gray value	Cluster number = 1, mean gray value < 84	832
	White	Cluster number + mean gray value	Cluster number = 1, mean gray value > 170	688
	Black-white mixed	Cluster number	Cluster number = 2	1200
	Color-mixed	Cluster number	Cluster number ≥ 3	784
Size	Large	Area	≥ 390	1152
	Medium	Area	121–389	2416
	Small	Area	< 121	1072
Illumination	Low light	HSV V channel mean	< 0.39	1520
	Normal light	HSV V channel mean	0.39–0.74	1824
	High light	HSV V channel mean	≥ 0.75	1296

categorized into low, normal, and high illumination levels (brightness thresholds: 0.39, 0.75) to simulate different lighting conditions. Table 1 presents the specific method for dataset partitioning.

For annotation, the dataset employed Labelme for pixel-level instance segmentation following the COCO format. Ten percent of the images were randomly selected for manual verification, and the average IoU between two rounds of annotation exceeded 0.95, confirming the accuracy and reliability of the labels.

This design ensures that the dataset exhibits high diversity and complexity across multiple dimensions, including color, shape, size, and illumination, while maintaining precise and reliable annotations. Therefore, it provides a solid foundation for evaluating different fusion strategies under complex input conditions and serves as a valuable benchmark for testing model generalization. Fig. 5 illustrates the dataset distribution.

3. Attention based adaptive feature Reorganization fusion algorithm

Feature-level fusion allows independent neural networks to extract features from different modalities and integrates them in a high-level semantic space, thereby more effectively capturing the complementary information between modalities. This fusion approach avoids the information redundancy caused by direct concatenation, enhances the feature representation capability of the extraction networks, and fully leverages the complementary features of different modalities. Various fusion strategies exist for this approach, typically employing bilinear pooling, weighted concatenation, attention mechanisms, and other techniques to learn the importance weights of features from different modalities, thereby improving the effectiveness of feature fusion.

The fusion of cross-modal information typically has a direct impact on model performance. To obtain more comprehensive and discriminative fused features, as illustrated in Fig. 6, this study integrates multi-level features from two perspectives using color and normal modalities as examples. First, features at different scales contain distinct information that can complement each other. Therefore, a multi-scale progressive fusion strategy is adopted to integrate single-modality features from coarse to fine. Second, for multimodal features, this study enhances the RGB and normal features separately instead of fusing them at an early stage. This strategy reduces interference between different modalities and preserves the effective information of each modality. Ultimately, multimodal feature fusion combines the features from both modalities to facilitate the identification of rock fragments.

3.1. AFRNet

This paper proposes a network named AFRNet to achieve feature-

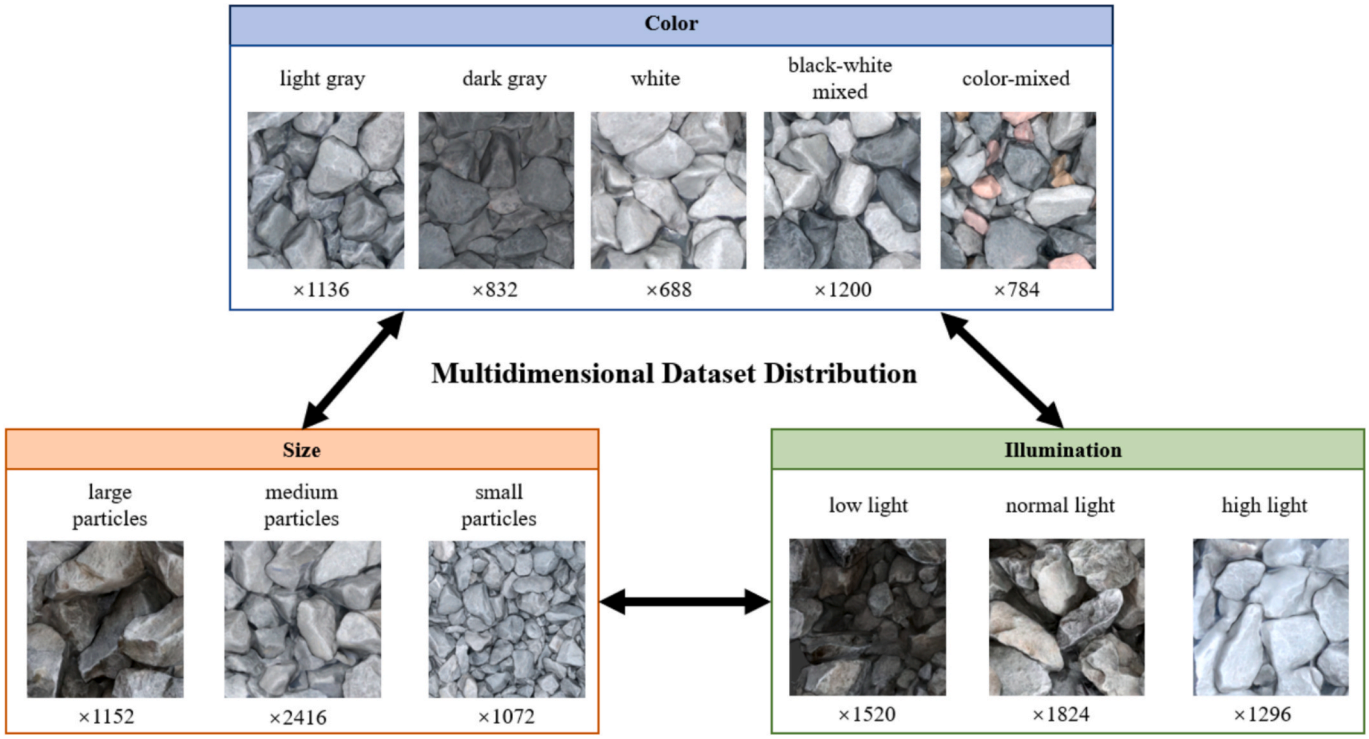


Fig. 5. Distribution of the Dataset.

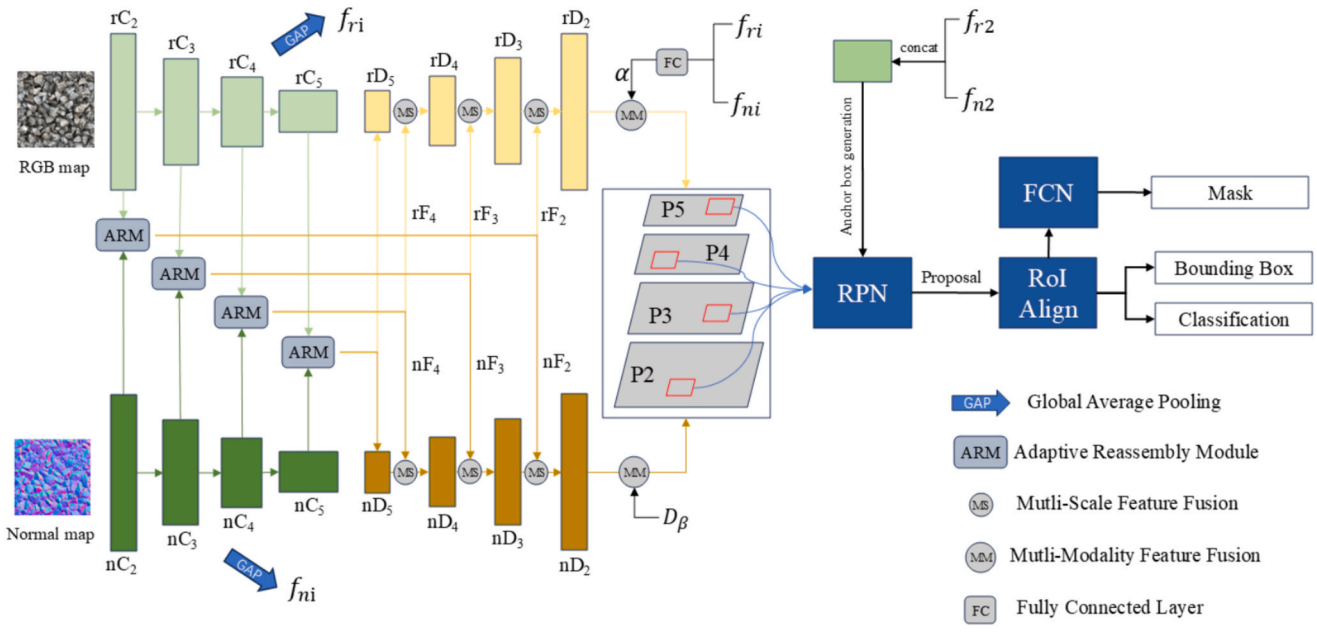


Fig. 6. Architecture of the AFRNet Model.

level fusion. The network integrates two independent ResNet backbones [48], employs an Auto Recombination Model (ARM) to enable interaction between features of different modalities, and utilizes Scale Progressive Multi-Model Fusion (SPMMF) to facilitate interaction among features at different scales. Additionally, the network incorporates a Region Proposal Network (RPN) [49], a RoI Align module, and a Fully Convolutional Network (FCN) [50].

The backbone networks are based on ResNet50, with the two backbones independently performing convolutions at the same channel dimension to ensure stable encoder operation and enable extraction of deeper features. Features derived from normal maps show distinct advantages in recognizing particle edges, while features from color images assist in target localization. Therefore, ARM is designed to adaptively process local features from different modalities to achieve complementary feature integration.

As convolutional layers deepen, low-level features provide richer detail information, such as boundaries, textures, and spatial structures, but are more susceptible to background noise. In contrast, high-level features contain stronger semantic information that aids target localization and noise suppression. In the decoder part, a Multi-Scale Feature Fusion (MSFF) module is designed to manage contextual information by adaptively focusing on both global and local features of particles, thereby improving boundary recognition and enhancing detection accuracy. The RPN generates candidate regions of interest (RoIs) for subsequent detection and recognition. The RoI Align module outputs multiple bounding boxes, which are resized to fixed dimensions by the Spatial Pyramid Pooling (SPP) [51] network and then fed into fully connected layers. Finally, a regression model is used to predict bounding boxes. The FCN outputs a binary mask for each RoI to achieve particle segmentation.

3.2. Reliability evaluation of bimodal priors

In the proposed network, the backbone is fully symmetrical. We denote the feature representations from different channel dimensions of the RGB modality encoder as rC_2, rC_3, rC_4, and rC_5, and those from the normal map modality encoder as rN_2, rN_3, rN_4, and rN_5. These features are then paired and fed into the ARM for fusion. Previous works generally merge features from the RGB modality and other modalities indiscriminately, which can introduce noise when the auxiliary modality is unreliable. Fan et al. [52] and Chen et al. [53] evaluated the reliability of auxiliary modalities by either discarding low-quality modalities or modeling confidence scores to control the fusion process based on quality assessment. However, in our previous work, we found that especially for the normal map modality, the confidence response to rock fragment recognition may even surpass that of the RGB modality. Therefore, we performed a bidirectional quality evaluation by converting the involved modality images into grayscale. If the grayscale image segmented by thresholding is closer to the ground truth (GT), the modality is deemed more reliable.

We designed an algorithm to achieve fast and lightweight prior reliability evaluation of modalities. Specifically, we apply Otsu's adaptive thresholding to the grayscale images of different modalities to obtain binary masks M. Then, the GT encoded in run-length encoding (RLE) is converted into a binary image G. We conduct a preliminary pixel-level evaluation using the Intersection over Union (IoU) metric to calculate D_{iou}.

$$D_{iou} = \frac{|I \cap G|}{|I \cup G|} \tag{2}$$

Here, the absolute value denotes the region size. However, directly applying thresholding to grayscale images usually introduces considerable noise, which may result in a low D_{iou} value even when a modality is strongly correlated with the ground truth (GT). To address this, we introduce recall to balance the bias of using only the IoU metric.

$$D_{re} = \frac{|I \cap G|}{|G|} \tag{3}$$

Recall reflects the proportion of the correctly identified particle regions relative to the true particle regions, encouraging the thresholded segmentation to more comprehensively cover the actual particles. Finally, we use a weighted harmonic mean to combine these two metrics:

$$D_{\beta} = \left(1+\beta^2\right) \cdot \frac{D_{iou} \cdot D_{re}}{\left(\beta^2 \cdot D_{iou}\right) + D_{re}} \tag{4}$$

where β is a weight control coefficient. Considering the ubiquitous presence of noise, we set $\beta = 0.5$ to emphasize the completeness of the coverage area. The computation of D is performed prior to the input of multimodal images into AFRNet and is used at each layer of ARM to

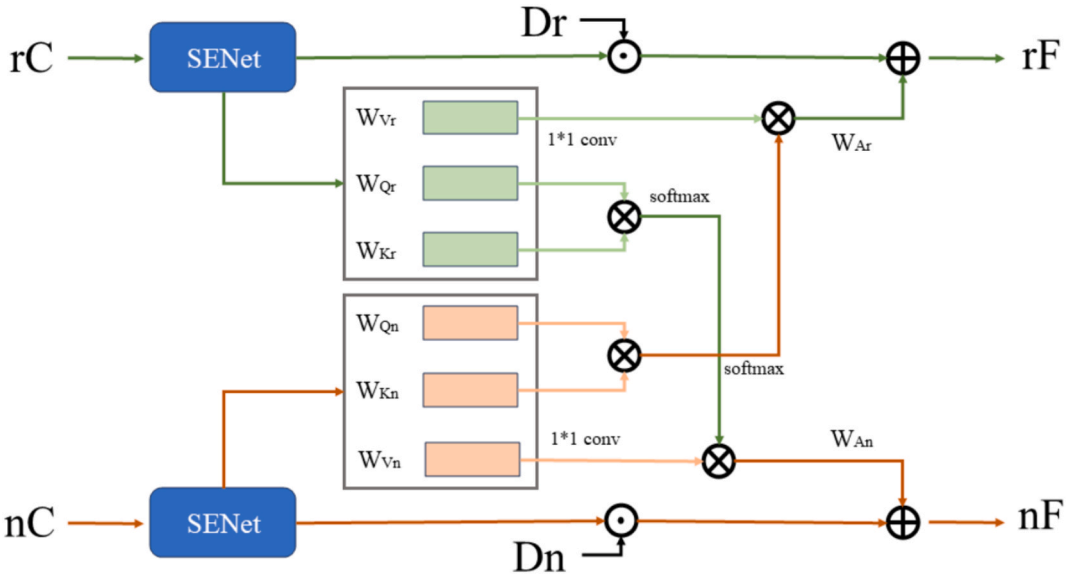


Fig.7. Structure of the ARM.

assist in determining the dominance of different modalities. The detailed design of ARM is elaborated in Section 3.3. Additionally, D is only available during training and does not participate in the inference process.

3.3. Adaptive Reassembly Module (ARM)

Considering the complementarity and inconsistency between different modalities, directly integrating cross-modal features using a fixed ratio often leads to suboptimal or even detrimental outcomes. Moreover, each modality may contain redundant information. To address these issues, we design an Adaptive Reassembly Module (ARM) integrated into a symmetric backbone network to enable interaction between different modalities, and we incorporate a dual-modal prior reliability assessment to prevent information contamination.

To reduce redundancy within single-modal features and to enhance the feature responses on particle regions, we apply channel attention using SENet separately to rC_i and nC_i. This process can be described as follows:

$$f = conv_1(f_{in}) \tag{5}$$

$$(W;B) = conv_1(f) \tag{6}$$

$$f_{out} = \delta(W \odot f + B) \tag{7}$$

Here, f_{in} denotes the input feature from either the RGB or normal branch (i.e., rC_i or nC_i); $conv_i$ (where $i = 1,2$) denotes convolutional layers; "$\cdot$" represents element-wise multiplication; δ is the ReLU activation function; and f_{out} denotes the refined RGB/normal feature at each stage. The number of channels of the refined features is unified to 256 dimensions.

In addition, to capture the dependency between cross-modal features, we introduce two cross-attention modules. As illustrated in Fig. 7, W_{Ar} utilizes normal information to generate geometric attention weights for RGB features, since normal maps typically offer richer edge cues to the RGB branch. Similarly, W_{An} uses RGB information to generate color attention weights for the normal features, as RGB features can provide global localization of target particles, thereby achieving cross-modal complementarity.

Specifically, we first apply a 1×1 convolution to nC_i, projecting it to W_Q and W_K, and project rC_i to W_v, where C, H, and W denote the number of channels, height, and width of W_v, respectively. C_1 is set to $C/8$ to reduce computational complexity. The enhanced feature is computed as follows:

$$f_A = softmax(W_Q \otimes W_K) \tag{8}$$

$$W_A = W_V \otimes f_A \tag{9}$$

The softmax operation is applied across the columns of W_A, and $\otimes$ denotes matrix multiplication. The enhanced feature rF is then reshaped into $C \times H \times W$. The other sub-module, W_{An} is symmetric to W_{Ar}.

Table 2
Key Parameters of the ARM Module.

Module	Input	Operation	Parameters	Output Size	Description
Channel Attention (SENet)	rC / nC	3 × 3 Convolution	Conv(3 × 3)	C × H × W	Extracts single-modal features
	–	Global Average Pooling	GAP	C × 1 × 1	Aggregates global channel information
	–	Fully Connected Layer	FC(256 → C), ReLU	C × 1 × 1	Learns channel weights to suppress redundancy
	f_in	1 × 1 Convolution	Eq.(7)	256 × H × W	Unifies channel dimension to 256 and enhances particle-region response
Cross-Attention Module	rC, nC	1 × 1 Convolution	W_Q, W_K, W_V (C → C_1, C'=C/8)	H × W × C'	Reduces dimensionality for computational efficiency
	–	Correlation Computation	Eq.(8)	(H × W)×(H × W)	Computes inter-modal attention map
	–	Weighted Fusion	Eq.(9)	C × H × W	Generates enhanced cross-modal features
	Dr, Dn	Residual Connection	rF = Dr ⊕ W_Ar, nF = Dn ⊕ W_An	C × H × W	Produces fused dual-modal features
Output	rF, nF	–	–	C × H × W	Outputs cross-modally enhanced feature representations

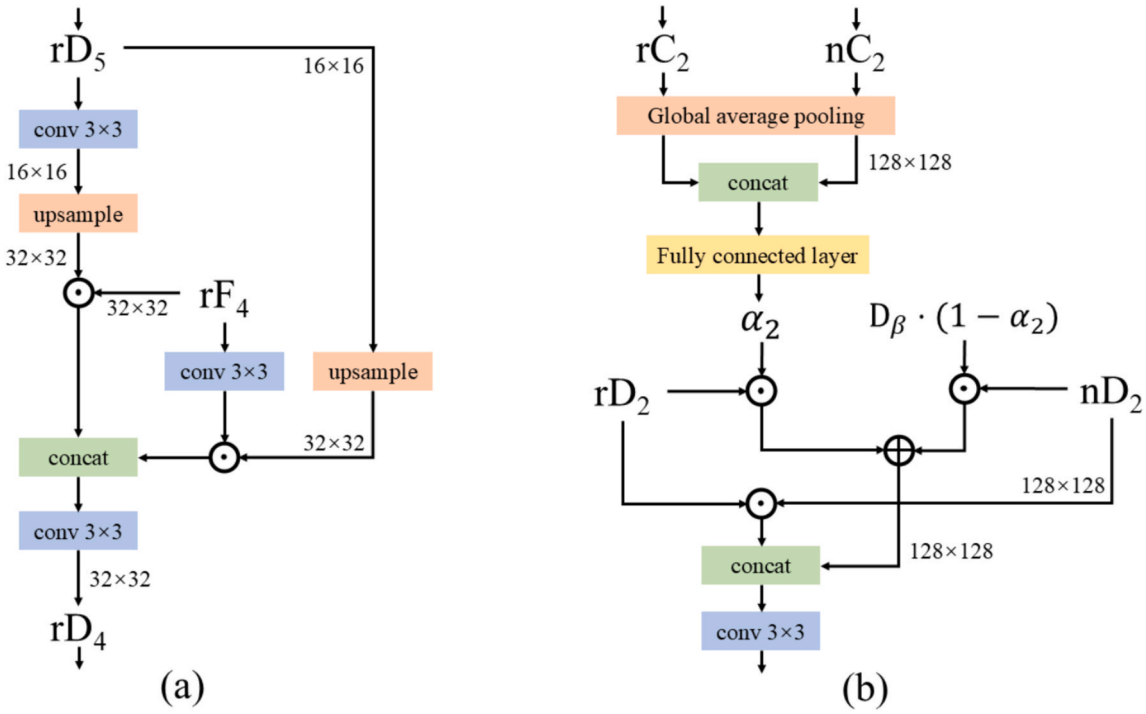


Fig. 8. Structure of SPMMF Module. (a) Multi-Scale Fusion Module. (b) Multi-Modality Fusion Module.

The introduction of these two attention mechanisms aims to achieve feature complementarity between the two modalities. The channel attention mechanism is designed to eliminate redundant information in each modality, preventing mutual interference caused by negative features. In contrast, the cross-attention mechanism is introduced to adaptively explore the dependencies between deep interactive features of different modalities, thereby optimizing the feature representations of both modalities. Table 2 shows the key node parameters of the ARM.

3.4. Scale progressive Multi-Model Fusion(SPMMF)

Due to the fact that different modalities tend to exhibit superior feature representations at different encoder depths—for instance, RGB and depth images often contribute more significantly at deeper layers, while normal maps are more effective in highlighting edge information at shallower layers—the complementarity among features at different scales is also a key factor in multimodal fusion. Therefore, feature fusion should not be limited to same-layer cross-modal interactions; it must also incorporate multi-scale feature fusion, which directly impacts the overall performance of the model.

To address this, we adopt a multi-scale progressive fusion strategy, which integrates features across modalities from small receptive fields to large ones. Subsequently, we enhance each modality's features using the ARM to determine the dominant modality at each layer, and ultimately aggregate the multimodal features to obtain the final representation. Fig. 8 shows the structure of the SPMMF Module.

1) Multi-Scale Feature Fusion(MSFF):

Low-level features provide rich detail such as edges, textures, and spatial structures, but are more susceptible to background noise. In contrast, high-level features contain more semantic information, which helps locate salient objects and suppress background interference.

Unlike previous approaches that commonly use concatenation or addition to merge low-level and high-level features, we adopt a more aggressive yet effective approach—element-wise multiplication. This operation enhances the response of salient objects while suppressing background noise.

Taking the fusion of high-level feature rD_5 and low-level feature rD_4 as an example, the multi-scale feature fusion process is illustrated as follows:

$$f_1 = \delta(upsample(conv(rD5)) \odot rF4) \tag{10}$$

$$f_2 = \delta(conv(rF4) \odot upsample(rD5)) \tag{11}$$

$$f_F = \delta(conv([f_1, f_2])) \tag{12}$$

Here, upsample denotes a bilinear interpolation-based upsampling operation, and $[\cdot, \cdot]$ represents the concatenation operation. The fusion result f_F is then used as the high-level feature input for the subsequent stage of fusion.

2) Multi-Modality Feature Fusion(MMFF):

In the process of multimodal feature fusion, in order to select the most useful and complementary information from both RGB and depth features, we learn a weighting vector α to balance the complementarity when fusing the two modalities. This weight is learned from the features of the dual-branch backbone networks and is aligned with the channel dimension; each element of the vector represents the importance of a specific channel. A broadcasting mechanism extends this weight across every pixel of the feature map, facilitating the identification of channels that are more informative for multimodal tasks.

To mitigate the potential interference caused by unreliable depth maps during fusion, we introduce an additional weight term, D, which controls the contribution ratio of depth information. This weight reflects the confidence in the reliability of the normal map.

Specifically, to fuse cross-modal features *rD* (RGB feature) and *nD* (normal feature), we design a weighted channel attention mechanism that automatically selects the most relevant channels. The fusion is computed as follows:

$$f_1 = \boldsymbol{\alpha_i} \odot rDi + D_\beta \cdot (1 - \boldsymbol{\alpha_i}) \odot nDi \tag{13}$$

$$f_2 = rDi \odot nDi \tag{14}$$

$$f_S = \delta(conv([f_1, f_2])) \tag{15}$$

Here, α represents the channel attention weight vector learned from both RGB and depth information, while D is the aforementioned gating

Table 3
Key Parameters of the SPMMF Module (a) Multi-Scale Feature Fusion Module (MSFM).

Layer Name	Input	Operation	Parameters	Output Size	Description
Conv1	rD_5	3 × 3 Convolution	Kernel = 3 × 3, Stride = 1, Padding = 1	16 × 16	Extracts high-level semantic features
Upsample1	Conv1 output	Bilinear Upsampling	Scale = 2	32 × 32	Upsamples to match lower-level feature size
Conv2	rF_4	3 × 3 Convolution	Kernel = 3 × 3, Stride = 1, Padding = 1	32 × 32	Extracts low-level detailed features
Fusion1	Upsample1 × Conv2	Element-wise Multiplication	–	32 × 32	Fuses high-level and low-level features (Eq.10–11)
Concat	f_1, f_2	Concatenation	–	32 × 32	Concatenates two fusion results along channel dimension
Conv3	Concat output	3 × 3 Convolution	Kernel = 3 × 3, Stride = 1, Padding = 1	32 × 32	Outputs fused feature rD_4

(b) Multi-Modality Feature Fusion Module (MMFM).

Layer Name	Input	Operation	Parameters	Output Size	Description
GAP	rC_2, nC_2	Global Average Pooling	–	1×1×C	Extracts global channel statistics for both modalities
Concat	GAP(rC_2), GAP (nC_2)	Concatenation	–	1×1×2C	Combines global features of RGB and normal modalities
FC	Concat output	Fully Connected Layer	Output channels = C	Generates channel attention weights α_2	Learns channel weights to suppress redundancy
Weighted Fusion	rD_2, nD_2	$\alpha_2 \odot rD_2 + D_\beta \cdot (1-\alpha_2) \odot nD_2$	Eq.(13)	128×128	Weighted fusion of RGB and normal features
Common Fusion	rD_2, nD_2	Element-wise Multiplication	Eq.(14)	128×128	Captures common responses between modalities
Concat	f_1, f_2	Concatenation	–	128×128	Combines complementary and shared features
Conv	Concat output	3×3 Convolution	Kernel=3×3, Stride=1, Padding=1	128×128	Produces final fused feature f_S

weight. Equation (15) captures the common response to particle targets, whereas Equation (14) fuses the two modalities by incorporating both channel selection and confidence control, taking into account both the complementarity and inconsistency between the modalities. Table 3 shows the key node parameters of the SPMMF.

3.5. Loss function

The loss function plays a crucial role in the training of deep learning models. It quantifies the discrepancy between the predicted outputs and ground truth labels, thereby guiding the optimization of the objective function, directing parameter updates, and providing a measurable performance indicator.

In this work, we adopt a multi-task loss function to supervise the training process. Specifically, a classification loss is employed to constrain target prediction accuracy, while a regression loss is designed to model the latent response capability of different modalities. As shown in Equations (16) and (17), the combined objective effectively balances semantic prediction and modality-specific representation learning [54,55].

$$L = \sum_{t=1}^{T} \alpha_t \left(L_t^{cls} + L_t^{reg} \right) + \beta L^{mask} + \gamma L^{sema} + \lambda L^{glbctx} \tag{16}$$

$$\beta = \sum_{t=1}^{T} \alpha_t \tag{17}$$

Specifically, L_{cls}, L_{reg}, L_{mask}, L_{sema}, and L_{glbctx} represent the classification loss, bounding box regression loss, mask prediction loss, semantic prediction loss, and global context loss, respectively. The total number of stages is denoted by T, which is set to 3 in our framework. The classification and regression losses (L_{cls} and L_{reg}) are applied across all three stages, and their sum constitutes the box loss, which follows the formulation in Cascade R-CNN [56]. The coefficients α_t and γ are used to balance the contributions from different stages and tasks.

The mask loss L_{mask} supervises the mask prediction and is only applied to valid samples using a mask head (MH). The corresponding weight βequals the sum of the loss weights across all stages. The semantic segmentation loss L_{sema} is formulated as a cross-entropy loss, while the global context loss L_{glbctx} is computed via binary cross-entropy for multi-label classification.

Table 4
Training Hyperparameters Used in the Experiments.

Hyperparameters	Value
Batch size	4
Learning rate	0.0001
epochs	300
Lr scheduler	T_max = 50, eta_min = 1e-6
Loss weights	1.0*cls_loss + 1.0*$bbox\_loss$ + 2.0*$mask\_loss$
Score thresh	0.05
NMS thresh	0.7

To optimize the proposed network more effectively, we incorporate auxiliary losses at four decoding stages. Specifically, for each stage (i.e., rD_i, i = 5,4,3,2. dD_i, i = 5,4,3,2), a 3 × 3 convolution layer compresses the feature maps to a single channel. These feature maps are then upsampled to the same resolution as the ground truth via bilinear interpolation and normalized to the [0, 1] range using a Sigmoid function. The total classification loss thus consists of both the main loss and the four auxiliary losses. We adopt: α = [1,0.5,0.25], β = 1.75, γ = 0.2, λ = 3.

4. Experiments and results

4.1. Training details

The experiments were conducted on a workstation equipped with 64 GB RAM, an Intel Xeon Gold 6226R CPU, and an NVIDIA RTX 3090 GPU. The deep learning framework used was PyTorch, and development was done in PyCharm. In this work, a dataset consisting of 400 sampled rock fragment boxes was constructed, with corresponding texture maps of size 2048 × 2048 in three modalities. These maps were cropped into 19,200 images of size 512 × 512 to form the dataset, where 13,440 images were used for training, 3,840 for validation, and 1,920 for testing. Annotation was performed using an interactive algorithm. Some of the hyperparameters used during training are listed in Table 4.

Additionally, in this study, since the training images include normal

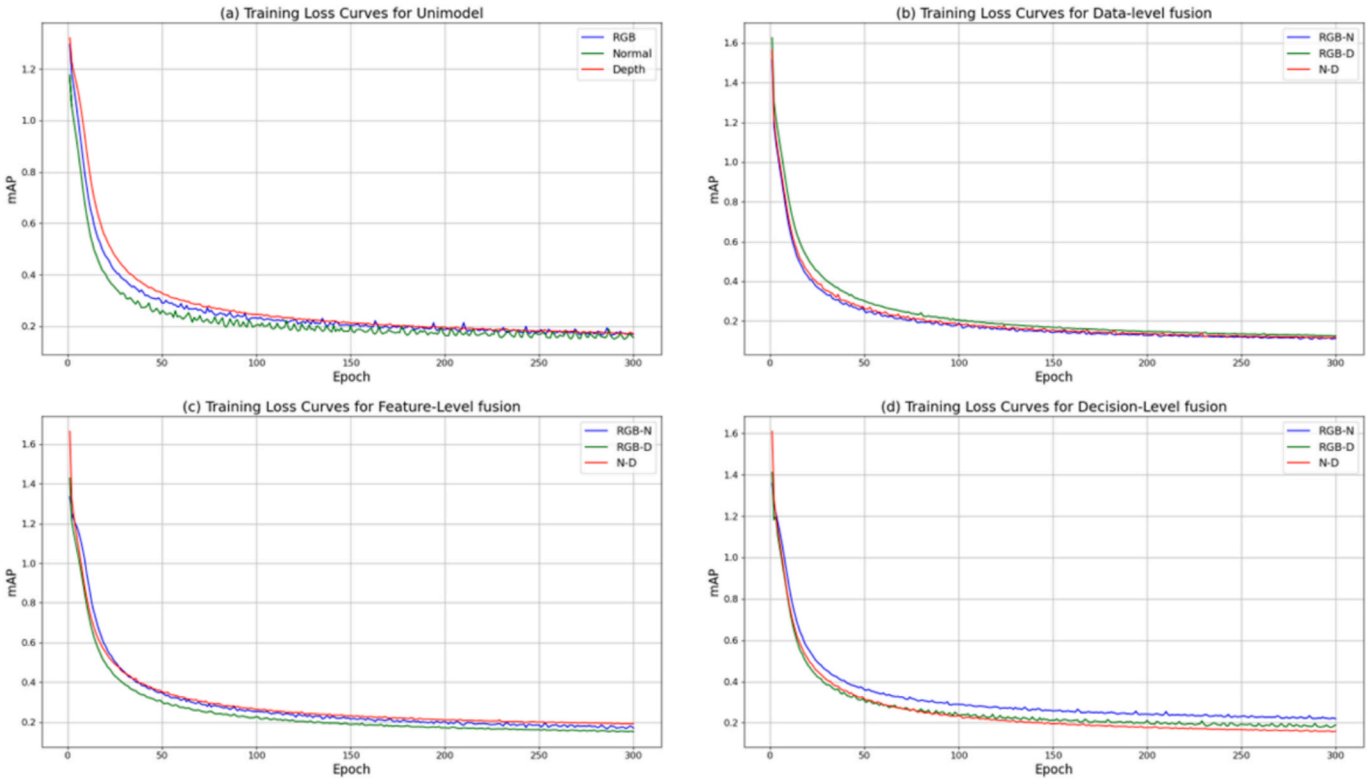


Fig.9. Training Loss Curves under Different Fusion Strategies.

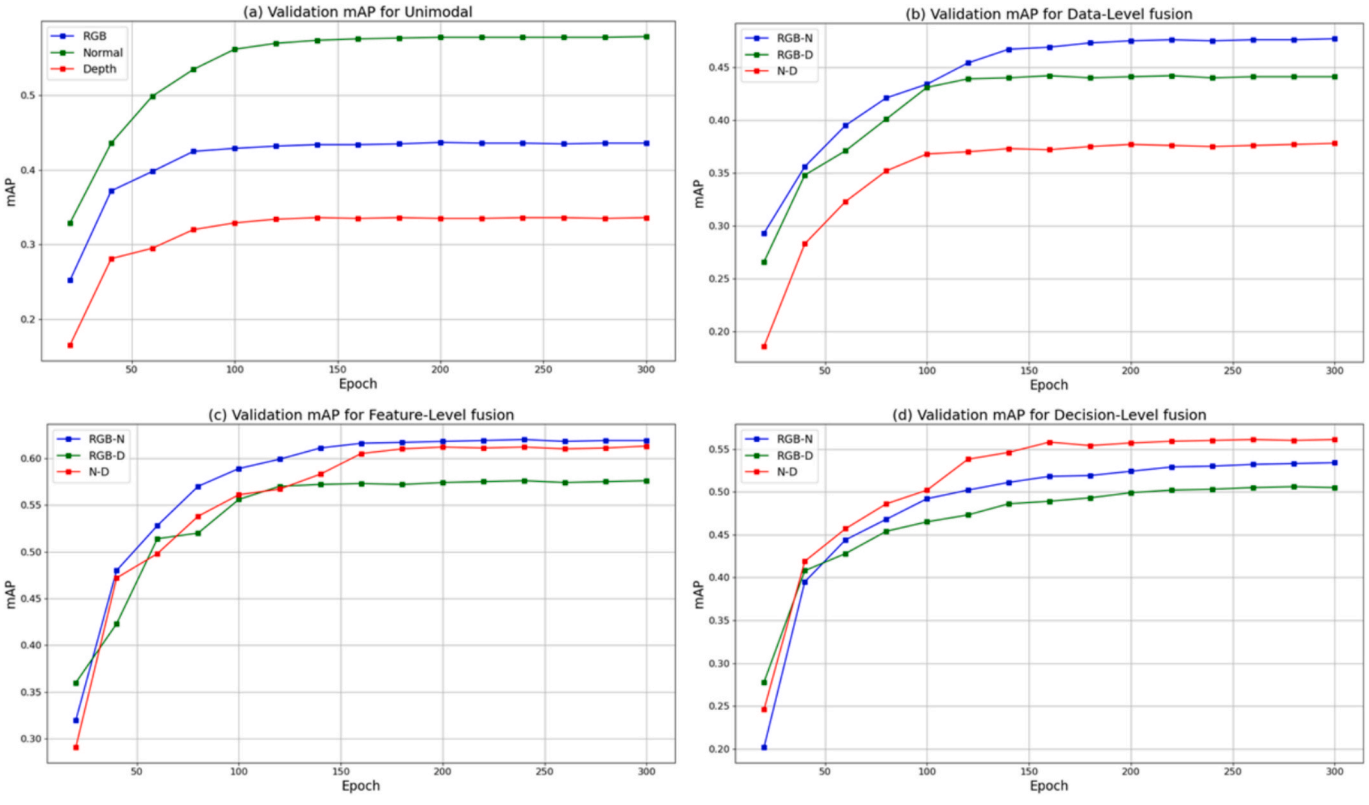


Fig.10. Validation mAP Curves under Different Fusion Strategies.

maps and depth maps, which differ significantly from natural images in pixel distribution and semantic structure, directly using models pre-trained on ImageNet may cause the early feature extraction layers to be ineffective. Therefore, for single-modality training of normal maps and depth maps, as well as data-level fusion training, a partial transfer learning strategy was adopted: only the first convolutional layer of the

Table 5
Map performance of different modalities and fusion strategies on the test set.

	modality	mAP_{50}	mAP_{75}	mAP
Unimodal	rgb	0.648	0.552	0.429
	normal	0.726	0.650	0.424
	depth	0.606	0.504	0.332
Data-levelfusion	rgb-n	0.672	0.567	0.382
	rgb-d	0.640	0.529	0.376
	n-d	0.572	0.465	0.291
Feature-levelfusion	rgb-n	**0.775**	**0.683**	**0.592**
	rgb-d	0.745	0.656	0.548
	n-d	0.761	0.647	0.576
Decision-level fusion	rgb-n	0.720	0.635	0.528
	rgb-d	0.709	0.602	0.489
	n-d	0.695	0.599	0.510
Mask2former	rgb	0.702	0.586	0.451
	rgb-n	0.697	0.524	0.437
SOLOv2	rgb	0.684	0.595	0.439
	rgb-n	0.698	0.571	0.442
YOLACT	rgb	0.619	0.500	0.408
	rgb-n	0.589	0.462	0.377
IAM	rgb-d	0.739	0.634	0.513
	rgb-n	0.730	0.625	0.524
SAM	rgb	0.726	0.602	0.512

Table 6
Comparison of Deployment Efficiency across Fusion Strategies and Competing Models.

Model	Params (M)	FLOPs (G)	FPS	Latency (ms)	Model Size (MB)
Mask R-CNN (R50)	44.5	132	17.9	55.9	172
Data-levelfusion	44.7	143	15.8	63.3	174
Feature-levelfusion	72.4	249	11.6	86.2	283
Decision-level fusion	85.3	276	9.2	108.7	365
Mask2former (Swin-T)	47.2	244	15.7	63.7	223
SOLOv2 (R50)	41.3	127	24.6	40.7	151
YOLACT (R50)	34.2	98	36.7	27.2	104
IAM (R50)	81.2	252	8.4	119.0	344
SAM (ViT-L)	308.1	1275	3.3	303.0	1228

backbone network was reinitialized, while the pretrained parameters of the remaining layers were retained and fine-tuned. The first convolutional layer mainly captures low-level textures and color features of the original input images, and its parameters cannot be directly reused for non-natural image types such as normal and depth maps, so retraining is necessary to adapt to these input modalities. Conversely, the higher layers have learned rich general semantic features (such as edges, local structures, and shape compositions) and possess good transferability. Retaining their pretrained weights can effectively improve model convergence speed and final performance. For the feature-level fusion scheme, due to the complexity of the fusion strategy and to avoid inconsistent convergence speeds among different modules during training, a random initialization approach was adopted.

4.2. Training and prediction results of each modality

The loss curves during the training process of various fusion strategies and different modalities are shown in Fig. 9. It can be observed that all models experience a rapid decrease in loss at the beginning of training, followed by gradual stabilization, consistent with the classical loss convergence trend. This indicates that each model has been sufficiently trained and has learned the features corresponding to its modality.

To improve training efficiency and reduce memory consumption, all models were validated once every 20 epochs. The mAP curves on the validation set during training with different fusion strategies and modalities are shown in Fig. 10, further illustrating the performance of each model and modality. Each curve generally shows a rapid increase in mAP followed by gradual stabilization, indicating that the models have demonstrated preliminary generalization ability on the validation set without overfitting. Among them, the feature-level fusion scheme based on attention mechanisms achieved the best results in the RGB-N modality, followed by the N-D modality. The mAP results on the test set are shown in Table 5, which are generally consistent with the final mAP results on the validation set. Table 6 shows the model performance parameters.

5. Discussion

5.1. Accuracy comparison of models under different fusion strategies

Data-level fusion refers to the method of stacking different modal data channels and directly inputting them into the instance segmentation model for training. This approach can fully preserve the original information of different modal images, especially low-level information such as color, texture, and edges, providing rich information sources for low-level feature extraction in segmentation tasks.

Since the multi-modality data used in this study are obtained via viewpoint rendering, they naturally share the same spatial resolution, allowing the direct use of standard convolutional neural networks for training without the need for additional feature matching. This paper improves the traditional Mask R-CNN framework to adapt it for multi-channel input.

Decision-level fusion adopts an ensemble learning approach, where data from different modalities are separately input into independent segmentation models, and the prediction results of these models are then combined. This is a feature-agnostic fusion method. Detailed descriptions of both fusion methods are provided in the appendix.

From the training loss curves and validation mAP curves, it can be seen that although the single-modality models and data-level fusion models have relatively lower final mAP values, they converge faster and achieve lower final loss values. This is due to two main reasons:

(1) Partial transfer learning was used. Typically, pretrained models are used during training to improve convergence speed and prediction accuracy. However, since the training data in this study includes normal maps and depth maps, directly using pretrained models trained on natural images may cause the shallow feature extraction ability to fail due to differences in pixel value distributions. Therefore, only the first convolutional layer of the backbone network was re-initialized, while the pretrained parameters of the other layers were retained and fine-tuned. Compared to training from random initialization, this leads to faster convergence.

(2) The models have relatively fewer parameters. Since the feature-level fusion models and decision-level fusion models use different feature extractors for different modalities, they have more parameters to update during training. By contrast, single-modality models and data-level fusion models use only one feature extractor, resulting in fewer parameters being updated during backpropagation and thus easier convergence. Additionally, single-modality datasets contain a simpler feature set, so fewer training iterations are needed for adequate model training.

Comparing the performance of the three modalities under different fusion strategies, it is found that in single-modality tasks, the normal map modality outperforms both the RGB and depth modalities on both validation and test sets. Among the two feature-level fusion models and the decision-level fusion model, although the rgb-n modality and n-d modality show different tendencies, both outperform the rgb-d modality. These three models share the characteristic of using

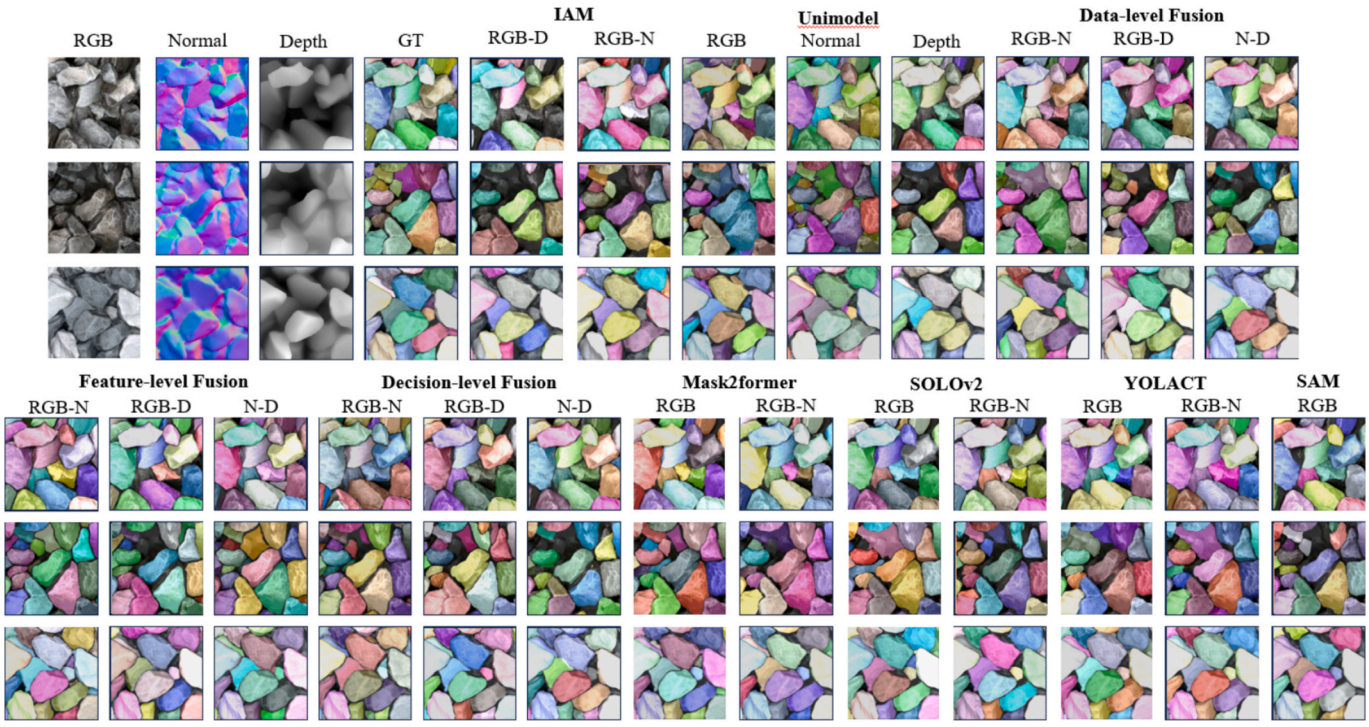


Fig.11. Qualitative Comparison of Prediction Results under Different Fusion Strategies.

separate feature extractors for different modalities, indicating that the features provided by the normal map modality significantly benefit instance segmentation and clearly outperform models without normal maps. Although the n-d modality performs worse than rgb-d in early fusion models, this is likely due to the channel stacking operation in early fusion disturbing the original feature distribution of images, so it does not directly reflect the advantages and disadvantages of each modality. Therefore, more attention should be paid to the performance of single modalities and other fusion approaches.

To enable a more comprehensive comparison, we additionally trained three representative instance segmentation models—Mask2-Former, SOLOv2, and YOLACT—using only RGB images. Furthermore, to keep the network architectures unchanged, we adopted an early fusion strategy by stacking RGB images and normal maps along the channel dimension and retrained the three models. The results show that early fusion brings a slight improvement to SOLOv2, while causing a slight performance drop in Mask2Former and YOLACT. However, none of these models outperform the proposed feature-level fusion approach. Although these algorithms are already highly mature, the absence of geometric information provided by normal maps limits their performance, resulting in inferior accuracy compared with AFRNet. In addition, simple early fusion without modality-specific filtering cannot effectively suppress inter-modality interference, making it difficult to achieve stable performance gains.

We also compared the Intramodal Attention Mix (IAM) module under different modality combinations. IAM was originally designed for indoor instance segmentation, where objects typically exhibit large depth variations. In contrast, objects in rock fragmentation scenes usually have a relatively small depth range. The incorporation of the N modality slightly improves the mAP of the IAM module, indicating the unique advantages of normal maps in representing rock fragment features. However, since IAM was originally designed for RGB-D multimodal input and was not specifically adapted for normal map input as in AFRNet, it does not demonstrate a clear advantage when considering the combined mAP_{50} and mAP_{75} metrics.

Table 5 shows that although the adaptive feature-level fusion model and the attention-based feature-level fusion model perform differently, both significantly outperform the early fusion model, with the early fusion model's test performance even falling short of the single normal map modality. This is because early fusion assumes all modalities share consistent semantic structures and distribution features at the raw pixel level. However, in practice, different modalities differ greatly in physical meaning, information representation, and noise characteristics. Simple concatenation often prevents the model from learning suitable low-level features for each modality effectively and may introduce interference among modalities, resulting in performance worse than single-modality models. By contrast, feature-level fusion methods encode each modality independently first and then fuse features at mid-to-high semantic levels, better meeting the needs of heterogeneous multimodal information integration. This also indicates that the feature-level fusion algorithms proposed in this paper successfully preserve modality-specific information structures, perform weighted integration of multi-source information in more abstract representational spaces, suppress inter-modality interference, and improve prediction accuracy.

5.2. *Inference efficiency comparison of instance segmentation models*

Table 6 presents a comparison of various instance segmentation models in terms of deployment performance, with a particular focus on inference speed, which reflects how differences in architectural design affect efficiency.

First, YOLACT and SOLOv2, as lightweight instance segmentation models, demonstrate significant advantages in speed. YOLACT employs a parallel structure for instance segmentation, bypassing complex processes such as region proposal generation and mask prediction. As a result, it achieves the highest frame rate (36.7 FPS) and the lowest inference latency (27.2 ms) among all models. SOLOv2, based on a single-stage spatial classification approach, also eliminates the traditional proposal stage, resulting in a high inference speed of 24.6 FPS.

Among the models based on Mask R-CNN (R50), the data-level fusion method only stacks different modality channels at the input level, while retaining the standard backbone (ResNet) and FPN structure. Therefore, its computational overhead is only slightly higher than the original Mask R-CNN, and its inference speed remains comparable (15.8 FPS vs. 17.9

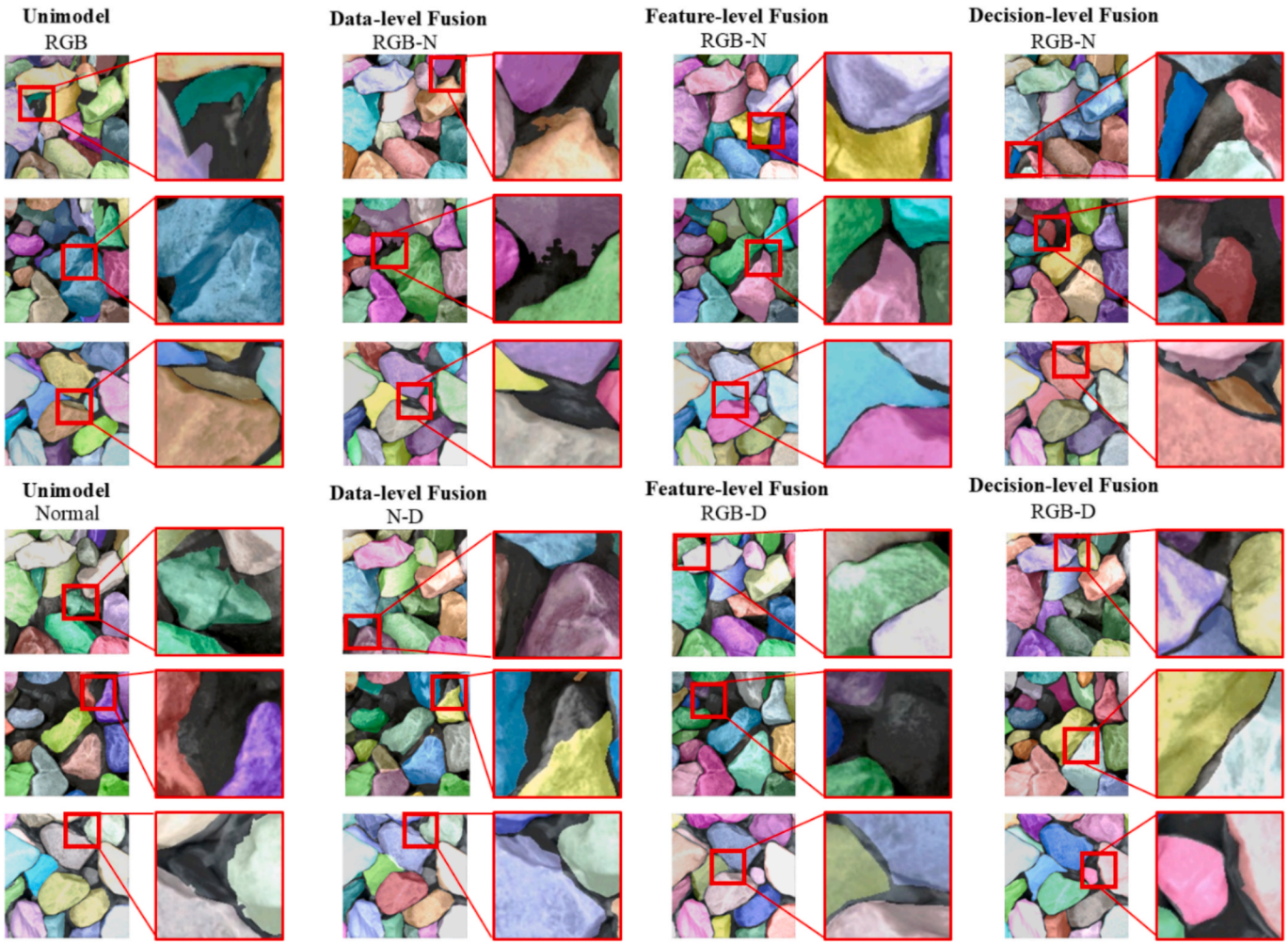


Fig.12. Detailed Visualization of Prediction Results under Different Fusion Strategies.

FPS).

The feature-level fusion method introduces dual-branch backbones, assigning a dedicated ResNet to each modality. This design significantly increases both the parameter count and FLOPs, leading to slower inference (11.6 FPS). However, to balance efficiency and fusion performance, the method shares a single FPN module for both modalities, partially mitigating the performance loss.

The decision-level fusion model consists of two entirely independent Mask R-CNN networks—each with its own backbone, FPN, and mask head—resulting in the largest parameter size (85.3 M), highest FLOPs (276G), and the slowest inference speed (9.2 FPS with 108.7 ms latency). This design emphasizes complete independence between modalities, making it suitable for applications that demand high accuracy but are less sensitive to inference time.

Mask2Former (Swin-T) represents a new generation of segmentation frameworks based on Transformer architectures. Although it incurs relatively high computational cost (244G FLOPs), its efficient feature aggregation enables a reasonable inference speed (15.7 FPS), outperforming several multimodal fusion models and striking a solid balance between accuracy and efficiency.

As a widely discussed universal segmentation framework in recent years, SAM (Segment Anything Model, ViT-L) employs a large-scale vision Transformer as its backbone and relies on extensive pre-training data to enable zero-shot or few-shot segmentation across diverse scenarios. However, this powerful representational capacity comes with extremely high computational costs. In the deployment evaluation of this study, SAM (ViT-L) exhibits a substantially larger number of parameters (308.1 M) and FLOPs (1275G) compared with other models, resulting in a significantly lower inference speed (only 3.3 FPS) and a high latency of 303.0 ms. Although SAM demonstrates certain advantages in cross-scenario generalization, its considerable computational burden limits its feasibility for direct deployment.

It is important to note that in practical rock engineering applications, such as outcrop fracture detection, particle size analysis, and 3D surface modeling, real-time processing is generally not a strict requirement. Instead, the accuracy and robustness of the results are of greater importance. Therefore, sacrificing inference speed in favor of improved predictive performance is a justified and acceptable trade-off. For instance, although feature-level and decision-level fusion methods introduce higher computational costs, they provide richer feature representations and better perception of complex textures and edges. This makes them more suitable for field environments with variable lighting and highly textured rock surfaces. Overall, the trade-off between deployment efficiency and segmentation accuracy should be determined by the specific task requirements, and multimodal fusion methods remain highly valuable for high-precision rock engineering applications.

5.3. *Particle recognition results*

Fig.11 illustrates the differences in particle recognition results among various models. it can be observed that the feature-level fusion models—based on both adaptive methods and attention mechanisms—have clear advantages in recognizing particle edges. these models can effectively allocate pixels in the shadowed connection areas

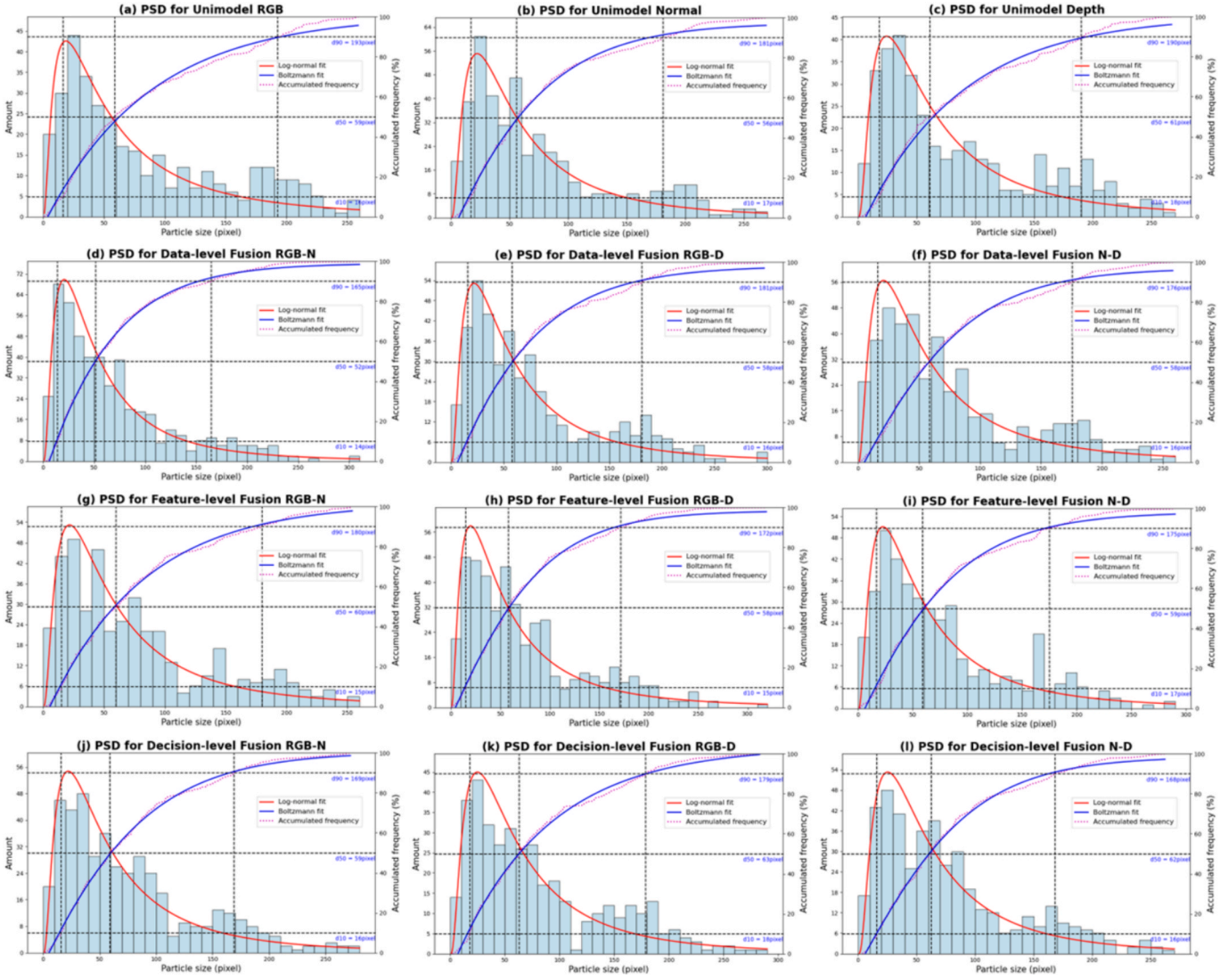


Fig.13. PSD Curve for Different Fusion Strategies.

between particles by leveraging shape and occlusion relationships, resulting in more natural segmentation outcomes.

For single-modality instance segmentation models, the normal map modality clearly outperforms the RGB and depth modalities. Normal maps inherently excel at representing edge features of particles, capturing pixel variations at particle corners more clearly. In contrast, RGB images are sometimes hindered by lighting conditions, which obscure edge shadows and reduce instance discrimination. Depth maps show advantages when identifying particle boundaries in stacked scenes, but their performance can degrade when height differences are small or when depth variation is caused by oblique viewing angles.

Although normal maps are strong at representing particle boundaries, they are less effective at localizing particles compared to depth or RGB images. As shown in Fig. 11, the instance segmentation model based on normal maps accurately segments the three particles in the upper-left region but misses the small particle in the lower-right corner, incorrectly classifying it as background.

In comparison, the feature-level fusion model based on the adaptive method not only retains the boundary-detection advantage of normal maps—accurately segmenting the three upper-left particles—but also incorporates the strengths of depth and RGB images in localizing particle instances. As a result, it successfully identifies the small particle in the lower-right area. This demonstrates that the proposed feature-level fusion algorithm effectively combines the strengths of multiple modalities while mitigating cross-modality interference.

Fig.12 presents representative details of segmentation results, providing a comparative analysis of different modality combinations and fusion strategies under complex conditions.

It can be observed that single-modality models show significant limitations when dealing with challenges such as low illumination, occlusion, or strong reflections in rock fragment images. These limitations are mainly reflected in incomplete segmentation of fragments in shadowed areas, adhesion between adjacent particles, and inaccurate or blurred boundary localization. Such phenomena indicate that single-modal input suffers from inherent blind spots in information acquisition, making it difficult to fully capture the complete geometry and boundary details of rock fragments.

After introducing multimodal information, the data-level fusion strategy—which directly stacks raw modal inputs—is theoretically expected to enhance representation capacity. However, the results suggest that this approach fails to effectively alleviate the aforementioned problems. In some cases, the fused model only segments the high-light regions of a particle, missing the overall shape, and even results in undetected fragments. This further highlights the insufficiency of data-level fusion in capturing true modality complementarity.

In contrast, the proposed feature-level fusion strategy demonstrates

significantly better performance in both boundary delineation and instance separation. Specifically, under the RGB-N modality combination, the model effectively integrates texture details from the RGB image and geometric cues from the normal map. The segmentation results exhibit clear fragment boundaries and complete instance separation, even under challenging conditions such as occlusion and highlight interference. Almost no false detections or missed detections are observed in these cases.

For the RGB-D modality, although the fusion model still performs well, missed detections occasionally occur in regions where the depth modality suffers from severe information loss. Nevertheless, it is important to note that the feature-level fusion has substantially mitigated this issue. Compared with using the depth modality alone—where structural loss and segmentation failure are more severe—the fusion model achieves significant improvements in accuracy and completeness.

This contrast further demonstrates that the introduction of the normal map as an auxiliary modality not only compensates for the structural perception deficiencies of RGB images but also plays a critical role in enabling effective multimodal feature integration. The normal map provides a structurally rich and complementary source of information that supports high-precision instance recognition and boundary localization.

Additionally, the decision-level fusion strategy, as a post-hoc ensemble method, achieves better overall accuracy than the single-modal and data-level fusion approaches. However, it still falls short in fully exploiting the complementary advantages across modalities. In complex regions, conflicts among predictions from individual models can occur, leading to suboptimal boundary integration and a lack of refinement in segmentation results.

In summary, these experimental results validate the effectiveness of incorporating normal maps as an informative modality for instance segmentation tasks. Furthermore, they provide qualitative evidence that the proposed feature-level fusion strategy offers stronger generalization capability and robustness in complex scenarios, thereby providing a reliable technical foundation for fine-grained intelligent segmentation of rock fragments.

5.4. Particle size and area statistics

The development of a multimodal fusion instance segmentation algorithm is aimed at facilitating the generation of Particle Size Distribution (PSD) for rock fragments. In this study, a 35 cm × 35 cm sampling box provides an accurate scale reference, allowing comparison between the segmentation results and standard criteria. For each rock fragment instance, the passing radius is characterized by the minimum rotated bounding rectangle, while the mass is represented by the upscaled area of the connected component. These measurements are used to plot particle grading curves and particle size distribution charts. The standard criterion (STCR) is established through meticulous manual annotation by experts using the WipFrag software, ensuring that the derived engineering parameters closely reflect real-world conditions. As shown in Fig. 13, ten images were randomly selected, and PSD analysis was conducted on all identified rock fragments within them. These correspond to cumulative results from ten sampling boxes of rock debris. Specifically:

1) The blue curves are used to fit the cumulative particle size distribution of rock fragments. In geotechnical applications involving excavation or blasting, rock fragments typically exhibit a distribution characterized by a scarcity of both fine and coarse particles, with a predominance of mid-sized fragments. Therefore, the Boltzmann model is employed for fitting:

2) The red curves are used to fit the frequency distribution of rock fragment sizes. Since particle size distribution is influenced by multiple multiplicative factors such as diagenesis, weathering, and fragmentation, the resulting particle sizes tend to approximate a normal distribution on a logarithmic scale. Therefore, a log-normal distribution model is adopted for fitting:

Table 7
Characteristic indicators for Different Fusion Strategies (mm).

	modality	d_{10}	d_{50}	d_{90}	d_{max}
Unimodal	rgb	±1.24	±3.57	±6.17	±2.43
	normal	±0.99	±4.81	±6.24	±5.36
	depth	±1.83	±4.76	±4.85	±1.69
Data-levelfusion	rgb-n	±1.26	±5.19	±6.33	±11.27
	rgb-d	±1.07	±4.56	±5.69	±6.34
	n-d	±2.27	±5.13	±5.94	±5.81
Feature-levelfusion	rgb-n	**±0.38**	**±2.39**	**±4.64**	**±1.04**
	rgb-d	**±0.36**	±3.55	±4.72	±1.58
	n-d	±0.43	**±2.86**	±5.67	±3.26
Decision-level fusion	rgb-n	±0.89	±4.04	**±4.62**	**±1.38**
	rgb-d	±1.28	±3.24	±4.98	±4.24
	n-d	±1.02	±4.51	±5.11	±7.52
Ground truth		3.92	12.14	32.08	53.89

3) The d_{10}, d_{50}, and d_{90} values annotated in Fig. 13 reflect the particle size characteristics of the corresponding rock fragments. In the field of geotechnical engineering, these parameters are of critical significance for understanding the physical and mechanical properties, permeability, and stability of geomaterials. Particle size and area statistics were computed in pixel units. Given the strict correspondence between the image dimensions and the sampling box size, the scale conversion factor between pixels and real-world length units is determined to be 1 pixel: 0.171 mm.

To thoroughly compare the results of particle size analysis and minimize the influence of rock fragment occlusion on the statistics, each sampling is subjected to ten rounds of shaking and image capture. All ten images are manually annotated, and the relative deviation in particle size analysis results is calculated. The cumulative statistical parameters from the ten annotations are then used as a reference. Table 7 presents a comparison of the parameters obtained from the algorithm and the STCR for all sampling boxes in the test set.

As shown in Table 7, the particle size parameters calculated by the STCR and the proposed method are generally consistent. This indicates that the PSD results obtained through the multimodal approach are applicable to real-world field analysis, further demonstrating the engineering applicability of the algorithm. Table 7 presents the relative errors of key particle size analysis metrics (d_{10}, d_{50}, d_{90}, and d_{max}) obtained from different instance segmentation models using various fusion strategies and modality combinations. The results demonstrate that both the fusion method and the choice of modalities significantly influence the accuracy of particle size measurements.

Among unimodal models, the RGB modality performs relatively well on d_{10} and d_{max}, while the normal modality shows larger errors for medium-to-large particles (d_{50} and d_{90}), particularly for d_{max} (±5.36 mm), indicating instability in detecting large fragments. The depth modality achieves the lowest error on d_{90} (±4.85 mm), suggesting its advantage in recognizing larger particles.

In the case of data-level fusion, where different modality channels are stacked directly at the input, the approach retains raw pixel-level information but lacks deep semantic interaction across modalities. This limitation results in notable segmentation inaccuracies. For instance, the rgb-n combination yields a large error on d_{max} (±11.27 mm), which is the highest among all methods. Further analysis shows that this is primarily due to the model mistakenly merging a large particle with surrounding smaller particles during segmentation, thus overestimating its size. In contrast, other fusion strategies are more stable in recognizing large particles, which typically have clearer edges and more distinct textures in images. This highlights the inefficiency of simple early fusion and the importance of deeper semantic integration.

Decision-level fusion, which processes each modality independently and combines results post-prediction, shows more robust performance than data-level fusion. However, it still lacks joint feature modeling and thus falls short in certain cases—for example, the n-d combination

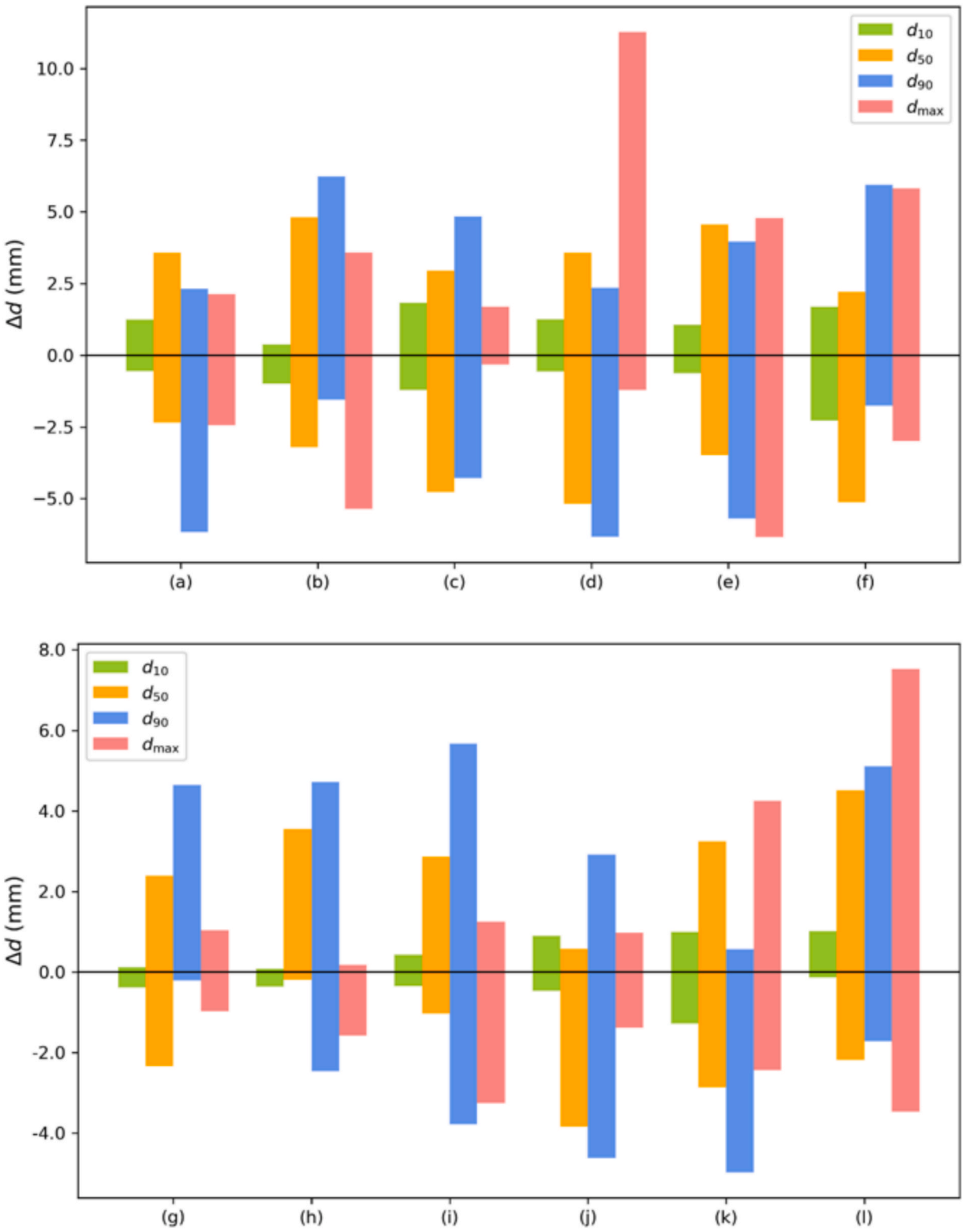


Fig.14. Fluctuations of Particle Analysis Metrics, where (a) to (l) represent the results of different fusion methods and modalities.

produces a relatively large d_{max} error (±7.52 mm), indicating difficulty in handling extreme particle sizes.

In comparison, feature-level fusion consistently achieves the best results across all modality combinations. By employing a dual-branch architecture with a shared FPN, this method allows independent deep feature extraction for each modality followed by effective semantic-level fusion. For the rgb-n combination, the errors for d_{10}, d_{50}, d_{90}, and d_{max} are only ± 0.38 mm, ±2.39 mm, ±4.64 mm, and ± 1.04 mm, respectively—significantly lower than other methods. Notably, the extremely low error on d_{max} demonstrates the model's exceptional capability in accurately segmenting large particles. This is particularly important in geotechnical engineering applications where maximum particle size often plays a critical role in material classification and performance evaluation.

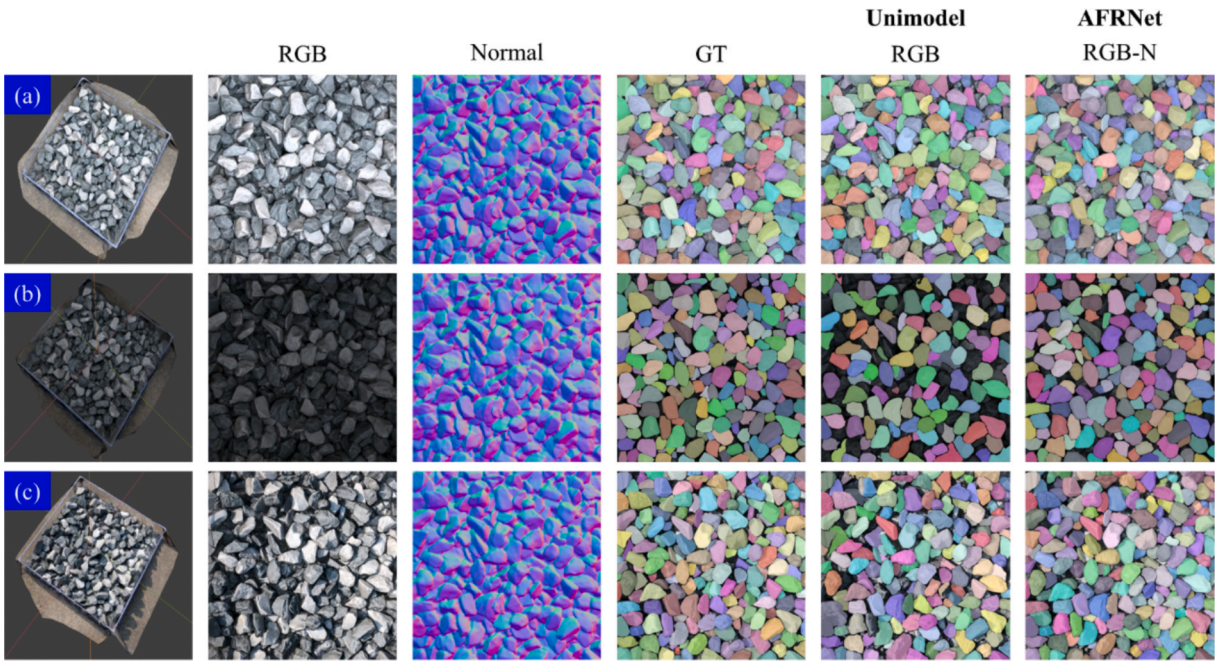


Fig. 15. Comparison of Recognition Results between Single-Modal RGB and Multi-Modal AFRNet under Different Lighting Conditions: (a) Normal Illumination, (b) Low Illumination, (c) Oblique Illumination with Shadows.

Table 8
Evaluation Metrics of Unimodel-RGB and Multi-Modal AFRNet-RGB-N under Different Lighting.

	Unimodel-RGB				AFRNet-RGB-N			
	mAP_{50}	mAP_{75}	mAP	AR	mAP_{50}	mAP_{75}	mAP	AR
(a)normal illumination	0.612	0.518	0.420	0.503	0.683	0.562	0.502	0.544
(b)low illumination	0.415	0.372	0.345	0.346	0.600	0.532	0.477	0.497
(c)oblique illumination	0.588	0.438	0.329	0.434	0.648	0.537	0.481	0.526

Moreover, a general trend across all models is that small particle recognition is less accurate than that of large particles, due to occlusion and blurred boundaries. However, in particle size analysis, small particles primarily influence d_{10}, which is typically a low-value metric. As a result, even with modest segmentation accuracy, the absolute error remains small and has limited impact on the overall analysis. This reflects a natural advantage of image-based methods for rock fragment analysis—namely, stable recognition of large particles and limited influence from small-particle errors on statistical results.

In summary, the feature-level fusion approach not only provides the most accurate segmentation across all particle size metrics but also excels in recognizing large fragments with minimal error. These advantages make it the most effective and practically valuable multimodal fusion strategy in this study. Fig. 14 further illustrates the maximum positive and negative deviations in particle size recognition results across different fusion strategies and modality combinations, offering additional insight into the variability and robustness of each approach.

5.5. Evaluation of AFRNet's illumination robustness enhanced by normal maps

To evaluate the stability improvement of AFRNet under different lighting conditions after integrating normal maps, we simulated three lighting scenarios in the laboratory: normal illumination, low illumination, and oblique illumination with shadows. Following the procedure described in Section 2.3, both RGB maps and normal maps were captured for each scenario. It can be observed that the color representation in the RGB maps varies significantly across the different lighting conditions, whereas the normal maps, although slightly degraded under low-light conditions, maintain a highly consistent overall appearance.

Based on these data, we compared the recognition results of the single-modal RGB model (Unimodel-RGB) and the dual-modal AFRNet-RGB-N under different lighting conditions. Fig. 15 illustrates the recognition results for the three lighting scenarios: (a) normal illumination, (b) low illumination, and (c) oblique illumination with shadows. Table 8 presents the corresponding evaluation metrics, including mAP and AR, verifying the advantage of dual-modal input in enhancing illumination robustness.

In the normal lighting condition, the recognition results of the single-modal RGB-based instance segmentation method are slightly worse than those of AFRNet, but the difference is not significant. However, in the low illumination scenario, all metrics show a substantial decrease, particularly the AR value, which drops from 0.503 to 0.346. As shown in Fig. 15, this is due to a large number of missed detections by Unimodel-RGB under low-light conditions. In contrast, although the recognition accuracy of AFRNet also decreases slightly, it remains significantly more stable than Unimodel-RGB. In the oblique illumination scenario, AFRNet's recognition results remain largely stable, whereas all metrics of Unimodel-RGB decline noticeably, with the mAP decrease being especially pronounced. This is because the shadows introduced by the oblique lighting interfere with the RGB modality's ability to capture particle edge features.

In summary, AFRNet with integrated normal maps demonstrates more stable recognition performance across various lighting conditions and significantly enhances robustness to illumination changes compared with the single-modal RGB model.

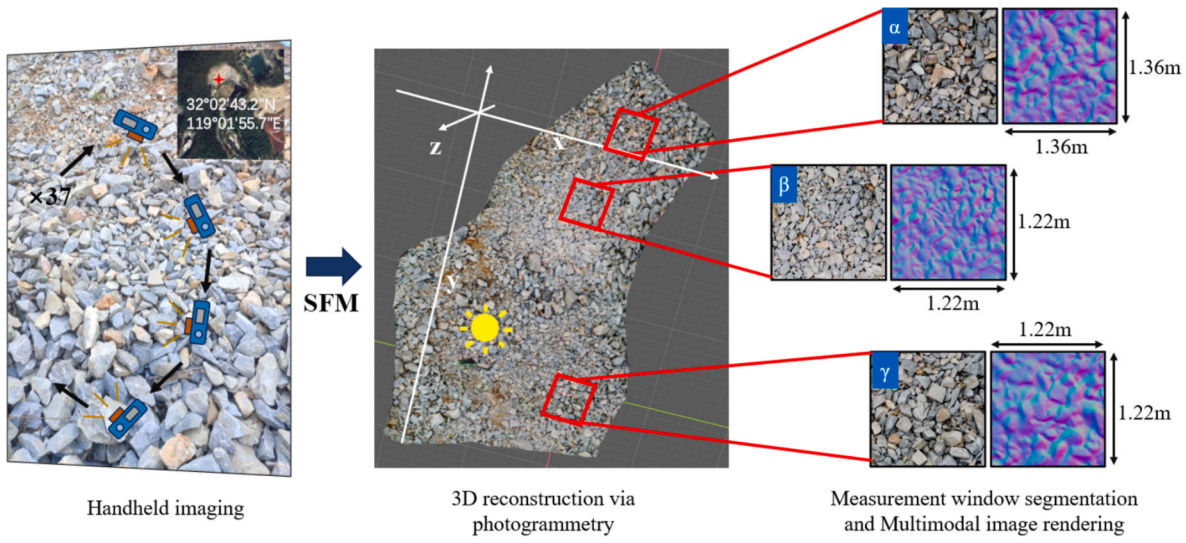


Fig.16. Workflow of the Case Study.

6. Real-World application: Open-Pit mine case study

To further validate the applicability and practical effectiveness of the proposed method in real-world mining scenarios, this study selects an abandoned open-pit mine in Nanjing as a case study and establishes a complete workflow for rock fragment recognition and particle size analysis. A handheld camera was used to capture photographs of the target area, which were then utilized for 3D reconstruction. Based on the reconstructed model, Physically Based Rendering (PBR) under consistent lighting conditions was applied to generate both color texture maps (RGB images) and normal maps, which serve as the input modalities for the subsequent instance segmentation task. Fig. 16 illustrates the workflow of the case study presented in this paper.

Multiple test windows were delineated on the generated orthophoto to create standardized testing samples, from which the corresponding color textures and normal maps were extracted. The proposed AFRNet was applied to perform instance segmentation on these multimodal images.

In the field of Mining and Blasting Engineering, the particle sizes of rock fragments at engineering sites typically span a wide range and exhibit substantial uncertainty, with some fragments reaching meter-scale dimensions. This makes on-site sieve analysis extremely costly or even infeasible. Consequently, particle size distribution (PSD) curves at engineering sites are generally not obtained through sieve testing. Instead, visual methods are employed—reference scales are placed on-site and photographs are taken, followed by manual annotation and algorithmic analysis to extract the PSD curves. Based on extensive prior research, the particle size distribution results obtained through manual visual annotation of rock images and subsequent scale conversion using a reference demonstrate acceptable deviations compared to those from sieve analysis [40,45,47,57]. Therefore, particle size analysis results obtained from image-based manual annotation are generally regarded as ground truth in practical engineering applications. This study converts the segmentation results from AFRNet into gradation analysis through scale conversion and compares them with other commonly used vision-based gradation analysis methods in engineering practice,

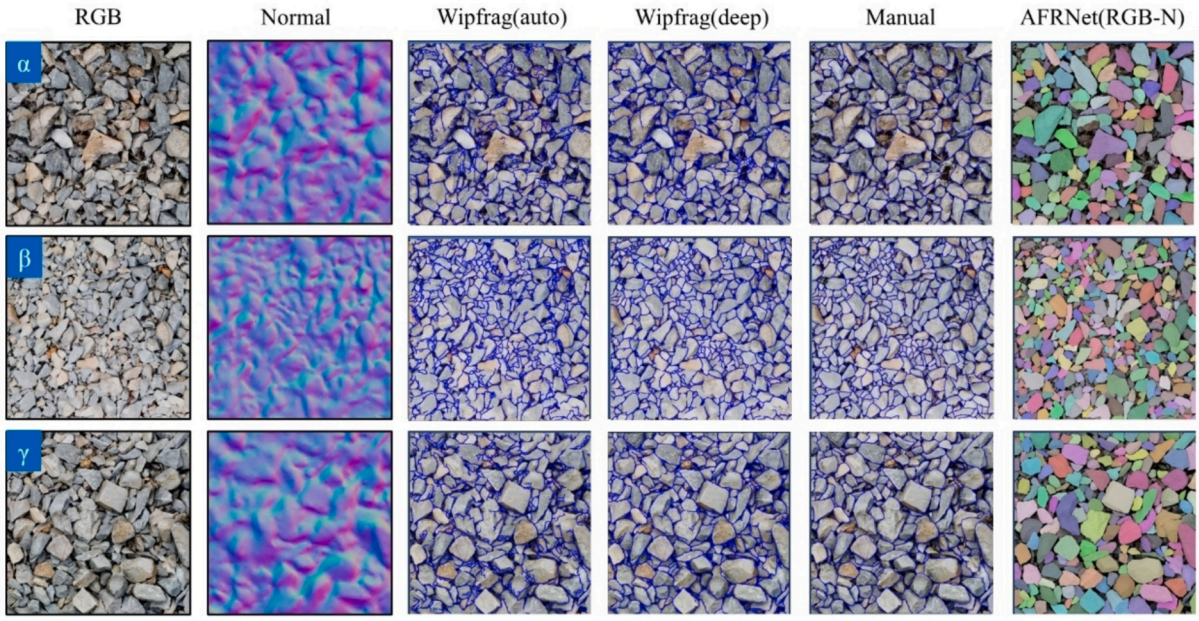


Fig.17. Comparison of Recognition Results for Mine Particles.

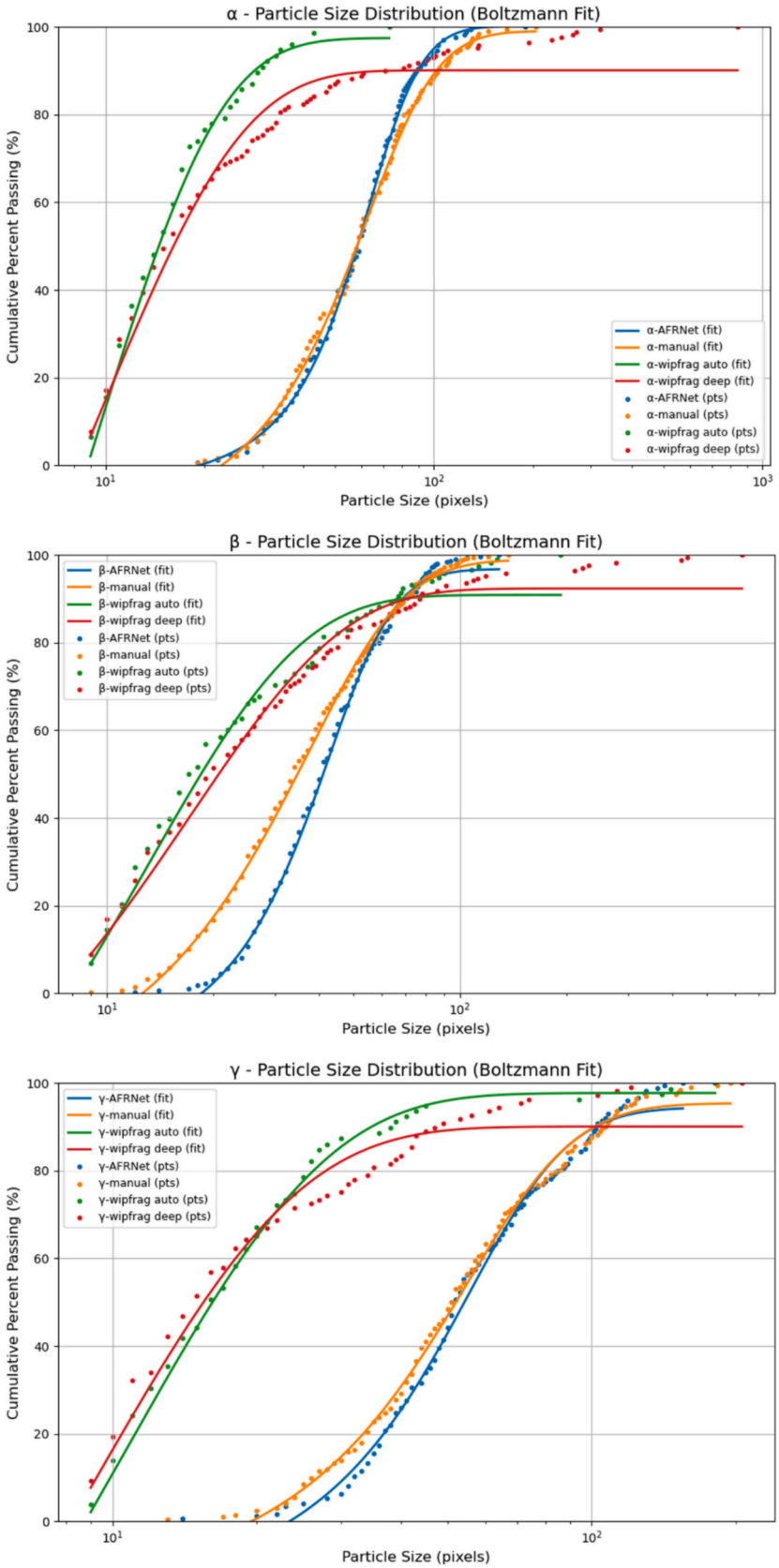


Fig.18. Analysis of Mine Particle Recognition Results.

Table 9
Particle Size Distribution Metrics for Different Segmentation Approaches (mm).

	Wipfrag(auto)			Wipfrag(deep)			AFRNet			Manual labeling		
	α	β	γ	α	β	γ	α	β	γ	α	β	γ
d_{10}	12.90	11.34	14.34	12.48	11.01	11.90	43.56	28.84	36.46	41.32	20.16	31.98
d_{50}	19.30	22.01	30.21	21.40	24.71	35.58	77.60	48.81	64.63	77.35	41.28	62.01
d_{90}	37.67	81.53	55.58	94.79	84.11	86.62	118.54	81.92	121.17	134.86	82.73	120.64

thereby validating the practical applicability of the proposed methodology. The WipFrag software is recognized as a standard tool. WipFrag supports fully manual annotation, incorporates built-in machine learning-based automatic segmentation algorithms, and also provides cloud-based deep learning inference for segmentation tasks. In practical engineering applications, PSD curves obtained using WipFrag can be directly used to guide construction. Fig. 17 presents a comparative visualization of segmentation results from AFRNet, WipFrag's built-in automatic segmentation, WipFrag's cloud-based deep learning segmentation, and manual annotations.

It is worth noting that WipFrag enforces a strict requirement for closed segmentation contours, which often leads to the erroneous identification of a large number of small particles when processing rock fragmentation images. This limitation is a common issue among commercial software used for rock fragment recognition in the field of geotechnical engineering. In contrast, the proposed AFRNet, although trained under laboratory conditions, demonstrates strong generalization capability when applied to on-site imagery. While the color characteristics of RGB images may vary under different environmental conditions, normal maps provide a more consistent and modality-invariant representation of particle geometry. As a result, the proposed multimodal segmentation method, enhanced by normal maps, exhibits robust and transferable performance across diverse scenarios. Although AFRNet may still miss a small number of fragments, its overall segmentation results are sufficiently stable and reliable for practical applications.

To further evaluate the engineering applicability of AFRNet's segmentation results in mining scenarios, particle size distribution (PSD) curves were plotted and fitted using the Boltzmann function. Key PSD metrics such as d_{10}, d_{50}, and d_{90} were then derived from the fitted equations for comparative analysis. Fig. 18 illustrates the PSD curves of segmentation results and manual annotations for three representative test windows, while Table 9 summarizes the corresponding particle size characteristics.

Table 9 presents a comparison of key particle size metrics (d_{10}, d_{50}, d_{90}) across three test windows (α, β, γ), obtained using four segmentation methods: WipFrag's built-in machine learning algorithm, WipFrag's cloud-based deep inference, AFRNet, and manual annotation. Overall, the particle size metrics extracted by AFRNet show a high degree of consistency with the manually labeled results, particularly in terms of d_{50} and d_{90}, indicating strong reliability in detecting medium- and large-sized fragments.

In contrast, the two WipFrag-based methods exhibit noticeable deviations in certain test windows. The primary reason lies in the strict boundary closure constraint imposed by both WipFrag's machine learning algorithm and its cloud-based deep inference model. While this constraint helps avoid merging adjacent fragments into a single object, it also leads to the generation of numerous redundant small connected regions, especially in areas with complex edges or heavy occlusion. This redundancy significantly affects the accuracy of particle size statistics, particularly d_{10}. By comparison, d_{90} is relatively less sensitive to such over-segmentation and remains reasonably close to the manual results in windows α and β.

The proposed AFRNet method leverages multimodal inputs of RGB images and normal maps, demonstrating robust boundary detection capabilities. Although it may still miss some very small particles, the overall particle size distribution curves are highly consistent with manual annotations. All metric deviations remain within 10 %, with most differences falling below 5 %, which confirms the engineering applicability and accuracy of AFRNet in complex mining environments.

7. Conclusion

This study focuses on the task of rock fragment extraction from images, introducing the normal map modality and developing a multimodal feature-level fusion instance segmentation method named AFRNet. The proposed approach enables precise analysis of rock fragments and demonstrates strong transferability across different scenarios. The main conclusions of this study are as follows:

Performance Improvement: In comparative experiments involving multiple modality combinations, feature fusion strategies, and advanced instance segmentation models, the RGB–N-based AFRNet demonstrates significant superiority. Quantitative results show that the model's mAP_{50}, mAP_{75}, and mAP outperform those of other modality combinations and state-of-the-art models. In visual evaluation, the segmentation results produced by RGB–N-based AFRNet more closely resemble real-world scenarios, surpassing other modality combinations and advanced instance segmentation models. Under varying lighting conditions, AFRNet-RGB-N exhibits markedly higher robustness to illumination variations compared with single-modal RGB models.

Robustness Enhancement: In experiments under different lighting conditions—including low illumination and shadow interference—the performance metrics (mAP and AR) and visual segmentation quality of instance segmentation models based on single-modal RGB drop significantly. In contrast, the RGB–N-based AFRNet maintains satisfactory segmentation results even in degraded environments. This indicates that the introduction of normal maps and the design of the feature fusion modules enhance the model's robustness against illumination disturbances.

Generalization Improvement: The AFRNet model trained in a laboratory environment is directly applied—without retraining—to a mining case in Nanjing, where its results are compared with those from traditional automated analysis using WipFrag software and network-based depth analysis. The generated particle size distributions and characteristic particle diameters (d_{10}, d_{50}, d_{90}) closely match the manually annotated results. The visualized outcomes further demonstrate excellent segmentation performance in real-world scenarios, validating the proposed method's transferability and engineering applicability.

Despite the proposed model's strong performance in recognition accuracy, robustness, and transferability, several limitations remain: (1) the feature-level fusion structure employs a dual-branch backbone, resulting in a large number of parameters and reduced training efficiency; (2) the method does not account for occlusion caused by fragment stacking, which may introduce bias in particle size statistics during practical applications. Future work may focus on network compression techniques such as model pruning and parameter sharing to develop more lightweight architectures and reduce computational costs. In addition, instance-level shape recovery algorithms and occlusion-region generation methods could be explored to improve the statistical reliability of particle-size analysis derived from segmentation results.

CRediT authorship contribution statement

Yulin Wang: Writing – original draft, Visualization, Validation,

Software, Methodology, Investigation, Formal analysis. **Xin Wang:** Validation, Software, Investigation, Formal analysis, Data curation. **Yudi Tang:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Xu Dai:** Visualization, Validation, Software, Data curation. **Jinming Dong:** Writing – review & editing, Visualization, Validation, Software. **Guangyao Si:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The grant support from the State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation(No. CDUT-PLC2024007) is acknowledged. The grant support from the China Scholarship Council (202208320010) is also acknowledged.

Appendix A. Data-Level and Decision-Level fusion architectures

In this study, the proposed model is specifically adapted for four-channel and six-channel inputs, as illustrated in Fig. A1. First, the preprocessed RGB images, depth maps, and normal maps are aligned and concatenated. To accommodate the increased number of input channels, the number of input channels in each convolutional kernel is adjusted accordingly. A deep neural network (DNN) is then employed to extract features from the multi-channel images, followed by a modified ResNet-50 or ResNet-101 backbone to generate feature maps. The ResNet architecture, composed of residual blocks and skip connections, allows efficient feature learning even in deep networks. Therefore, the use of a similar residual network structure in the proposed model contributes to improved system performance. Second, a Feature Pyramid Network (FPN) is constructed on top of the modified DNN to enhance multi-scale object segmentation. FPN improves segmentation accuracy for targets of various scales and strengthens the generalization ability of the model, without significantly increasing computational complexity. To implement the FPN, feature layers C_2 and C_3 (with spatial resolutions reduced by factors of 2 and 3, respectively) are extracted. In addition, feature layers C_4 and C_5, with downsampling factors of 4 and 5, are obtained from the residual blocks of the feature extractor to further enrich the pyramid structure. The resulting multi-scale feature layers are then fed into the subsequent network stages to boost segmentation performance. Third, a Region Proposal Network (RPN) is used to generate candidate object regions. As shown in Fig. A1, the upper branch of the RPN is responsible for classifying each anchor box as foreground or background, while the lower branch applies bounding box regression to refine the anchor locations. During training, the RPN optimizes both the classification loss and the bounding box regression loss. After processing, the RPN produces a set of candidate regions with improved localization accuracy. Finally, in the head of the model, a second-stage bounding box regression is performed along with object classification to obtain the final detection and segmentation results.

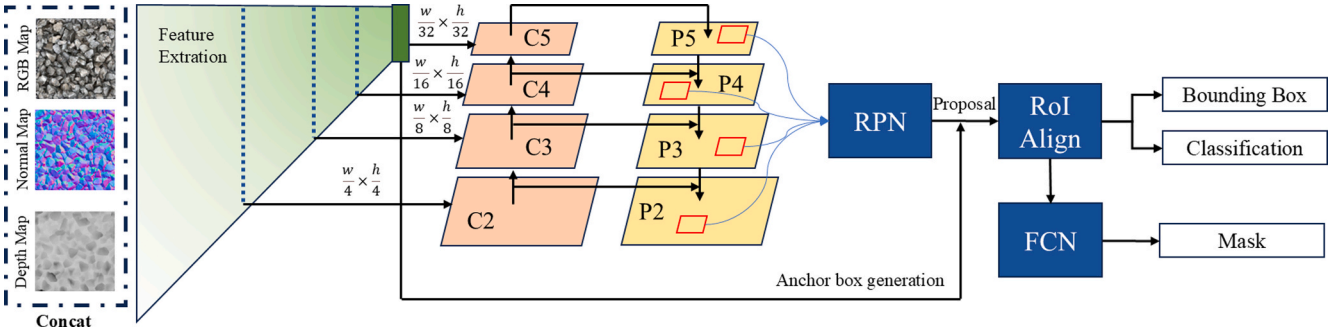


Fig. A1. Architecture of Data-Level Fusion Implementation

Decision-level fusion adopts the idea of ensemble learning, where data from different modalities are input into independent segmentation models, and then the prediction results of each model are merged. It is a fusion method that is not based on features. This fusion method can avoid feature interference between modalities, and uses different models for different modal data to construct the optimal architecture, thereby improving the overall performance of the model.

In this paper, a fully convolutional network is trained separately for fusion decision of recognition results from different modalities, assigning dynamic weights to each instance. As shown in Fig. A2, the IoU rule is first used for instance alignment. Masks judged to be the same instance are flattened into vectors. If an instance exists independently in different modalities, it is directly compared with the corresponding annotated mask and assigned a confidence weight of $a^*\ (M_{rgb} - i) + b^*\ (M_{depth} - j)$.

Then, the MLP input is $X = [Flatten(M_{RGB}),\ Flatten(M_{Depth})]$, and the result trained through three hidden layers is:

$$M_{fusion-k} = [\sigma(W_3 \bullet ReLU(W_2 \bullet ReLU(W_1 X + b_1) + b_2) + b_3)] \tag{A.1}$$

Where W_1, W_2, W_3, b_1, b_2, b_3, and σ are the linear parameters of the three hidden layers; ReLU is the nonlinear term. The final result is obtained by merging the independently weighted instances.

$$M_{fusion} = \sum \left[M_{fusion-k} + \widehat{a} \bullet M_{rgb-i} + \widehat{b} \bullet M_{normal-j}\right] \tag{A.2}$$

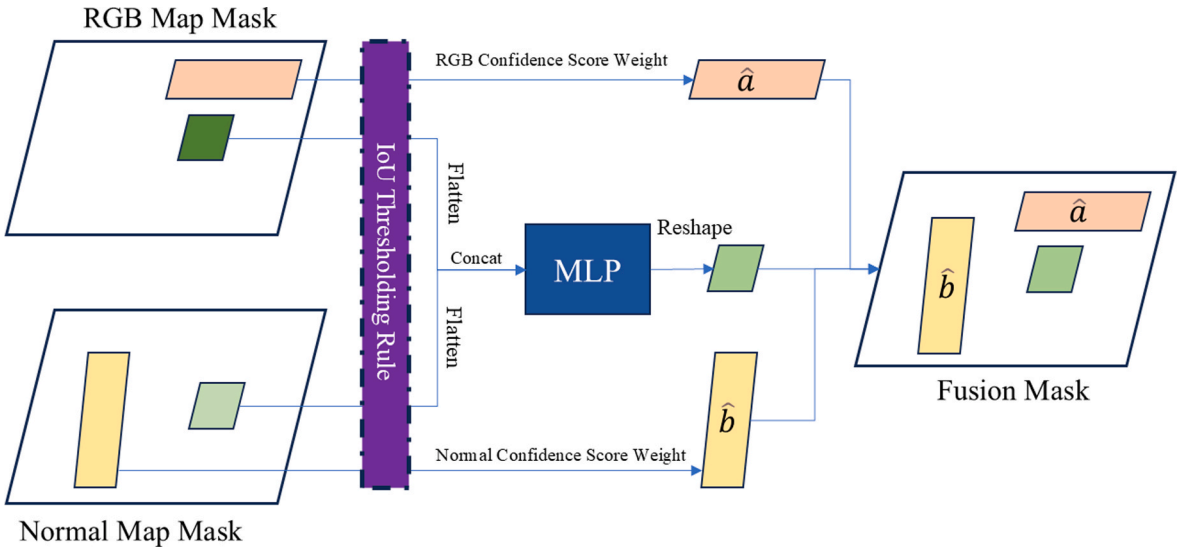


Fig. A2. Architecture of Decision-Level Fusion Implementation

Data availability

Data will be made available on request.

References

[1] S. Kumar, R.R. Jha, 40 Publications 458 Citations See Profile, (n.d.).
[2] T. Bamford, K. Esmaeili, A.P. Schoellig, A deep learning approach for rock fragmentation analysis, Int. J. Rock Mech. Min. Sci. 145 (2021) 104839, https://doi.org/10.1016/j.ijrmms.2021.104839.
[3] G. Huang, C. Qin, H. Wang, C. Liu, TBM rock fragmentation classification using an adaptive spot denoising and contour-texture decomposition attention-based method, Tunnelling and Underground Space Technology 161 (2025) 106498, https://doi.org/10.1016/j.tust.2025.106498.
[4] A. Rabbani, D.R. Kumar, Y. Fissha, N.P.G. Bhavani, S.K. Ahirwar, S. Sharma, B. K. Saraswat, H. Ikeda, T. Adachi, Optimization of an Artificial Neural Network Using Four Novel Metaheuristic Algorithms for the Prediction of Rock Fragmentation in Mine Blasting, J. Inst. Eng. India Ser. D (2024), https://doi.org/10.1007/s40033-024-00781-x.
[5] B. Cheng, I. Misra, A.G. Schwing, A. Kirillov, R. Girdhar, Masked-attention Mask Transformer for Universal Image Segmentation, (2022). doi: 10.48550/arXiv.2112.01527.
[6] X. Wang, R. Zhang, T. Kong, L. Li, C. Shen, SOLOv2: Dynamic and Fast Instance Segmentation, (2020). doi: 10.48550/arXiv.2003.10152.
[7] D. Bolya, C. Zhou, F. Xiao, Y.J. Lee, YOLACT: Real-time Instance Segmentation, (2019). doi: 10.48550/arXiv.1904.02689.
[8] A. Jung, S. Choi, J. Min, S. Hong, IAM: Enhancing RGB-D Instance Segmentation with New Benchmarks, (2025). doi: 10.48550/arXiv.2501.01685.
[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment Anything, (2023). doi: 10.48550/arXiv.2304.02643.
[10] Y. Tang, L. He, W. Lu, X. Huang, H. Wei, H. Xiao, A novel approach for fracture skeleton extraction from rock surface images, Int. J. Rock Mech. Min. Sci. 142 (2021) 104732, https://doi.org/10.1016/j.ijrmms.2021.104732.
[11] Z. Xiang, Z. Yu, W.-H. Kang, G. Si, J. Oh, I. Canbulat, Estimation of in-situ rock strength from borehole geophysical logs in Australian coal mine sites, Int. J. Coal Geol. 269 (2023) 104210, https://doi.org/10.1016/j.coal.2023.104210.
[12] R. Hu, G. Wang, Y. Wang, I. Canbulat, G. Si, From dynamic goaf formation to evolving spontaneous combustion and gas explosion hazard management, Process Saf. Environ. Prot. 198 (2025) 107131, https://doi.org/10.1016/j.psep.2025.107131.
[13] Y. Tang, L. He, H. Xiao, R. Wang, W. Lu, T. Xu, Fracture Extraction from Smooth Rock Surfaces Using Depth Image Segmentation, Rock Mech. Rock Eng. 54 (2021) 3873–3889, https://doi.org/10.1007/s00603-021-02481-4.
[14] Z. Xiang, W.-H. Kang, Y. Ji, G. Si, I. Canbulat, H. Lin, J. Oh, Estimation of in-situ horizontal stresses based on multiscale borehole breakout data via machine learning: model development, validation and application, Geophysical Journal International 242 (2025) ggaf144, https://doi.org/10.1093/gji/ggaf144.
[15] X. Wu, J. Dong, R. Hu, B. Pang, G. Si, CFD Modelling of Prevention and Mitigation of Coal Spontaneous Combustion in Longwall Goaf - A Comprehensive Review and Future Outlook, Arch Computat Methods Eng (2025), https://doi.org/10.1007/s11831-025-10420-7.
[16] K.A. Idowu, B.M. Olaleye, M.A. Saliu, Application of Split Desktop Image Analysis and Kuz-Ram Empirical Model for Evaluation of Blast Fragmentation Efficiency in a Typical Granite Quarry, GM 21 (2021) 45–52, https://doi.org/10.4314/gm.v21i1.5.
[17] M.O. Otokiti, B. Adebayo, Evaluation of Rock and Explosive Properties for Fragmentation Characterization: An Application of WipFrag, IJEATS 12 (2024) 13–32, https://doi.org/10.37745/ijeats.13/vol12n11332.
[18] H. Liu, M. Yao, X. Xiao, Y. Xiong, RockFormer: A U-Shaped Transformer Network for Martian Rock Segmentation, IEEE Trans. Geosci. Remote Sensing 61 (2023) 1–16, https://doi.org/10.1109/TGRS.2023.3235525.
[19] N. Saxena, R.J. Day-Stirrat, A. Hows, R. Hofmann, Application of deep learning for semantic segmentation of sandstone thin sections, Comput. Geosci. 152 (2021) 104778, https://doi.org/10.1016/j.cageo.2021.104778.
[20] F. Wang, Y. Zai, Image segmentation and flow prediction of digital rock with U-net network, Adv. Water Resour. 172 (2023) 104384, https://doi.org/10.1016/j.advwatres.2023.104384.
[21] J. Zhong, J. Zhu, J. Huyan, T. Ma, W. Zhang, Multi-scale feature fusion network for pixel-level pavement distress detection, Autom. Constr. 141 (2022) 104436, https://doi.org/10.1016/j.autcon.2022.104436.
[22] W. Qiao, Y. Zhao, Y. Xu, Y. Lei, Y. Wang, S. Yu, H. Li, Deep learning-based pixel-level rock fragment recognition during tunnel excavation using instance segmentation model, Tunn. Undergr. Space Technol. 115 (2021) 104072, https://doi.org/10.1016/j.tust.2021.104072.
[23] Z. Dai, H. Sun, L. Zhao, X. Qin, L. Wei, X. Liang, T. Chen, J. Jia, Intelligent Identification and Gradation Calculation of Aggregated Ore and Rock Particles Based on Instance Segmentation, Rock Mech. Rock Eng. (2025), https://doi.org/10.1007/s00603-025-04726-y.
[24] S. Du, W. Wang, R. Guo, R. Wang, S. Tang, AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Seattle, WA, USA, 2024: pp. 7608–7615. doi: 10.1109/CVPRW63382.2024.00756.
[25] J. Pan, S. Zhong, T. Yue, Y. Yin, Y. Tang, Multi-Task Foreground-Aware Network with Depth Completion for Enhanced RGB-D Fusion Object Detection Based on Transformer, Sensors 24 (2024) 2374, https://doi.org/10.3390/s24072374.
[26] X. Tang, S. Cen, Z. Deng, Z. Zhang, Y. Meng, J. Xie, C. Tang, W. Zhang, G. Zhao, Cascading attention enhancement network for RGB-D indoor scene segmentation, Comput. Vis. Image Underst. 259 (2025) 104411, https://doi.org/10.1016/j.cviu.2025.104411.
[27] W. Zhou, H. Zhang, W. Qiu, Differential Modal Multistage Adaptive Fusion Networks via Knowledge Distillation for RGB-D Mirror Segmentation, IEEE Trans. Big Data 11 (2025) 1959–1969, https://doi.org/10.1109/TBDATA.2024.3505057.
[28] L. Jiang, J. Zhang, B. Deng, Robust RGB-D Face Recognition Using Attribute-Aware Loss, (2019). doi: 10.48550/arXiv.1811.09847.
[29] Y. Fu, J. Fan, S. Xing, Z. Wang, F. Jing, M. Tan, Image Segmentation of Cabin Assembly Scene Based on Improved RGB-D Mask R-CNN, IEEE Trans. Instrum. Meas. 71 (2022) 1–12, https://doi.org/10.1109/TIM.2022.3145388.
[30] Y. Xiang, C. Xie, A. Mousavian, D. Fox, Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation, (n.d.).
[31] J. Ni, Z. Zhang, K. Shen, G. Tang, S.X. Yang, An improved deep network-based RGB-D semantic segmentation method for indoor scenes, Int. J. Mach. Learn. & Cyber. 15 (2024) 589–604, https://doi.org/10.1007/s13042-023-01927-1.
[32] S. Gupta, P. Arbeláez, R. Girshick, J. Malik, Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation,

Int J Comput Vis 112 (2015) 133–149, https://doi.org/10.1007/s11263-014-0777-6.
[33] R. Wang, W. Wan, Y. Wang, K. Di, A New RGB-D SLAM Method with Moving Object Detection for Dynamic Indoor Scenes, Remote Sens. (Basel) 11 (2019) 1143, https://doi.org/10.3390/rs11101143.
[34] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, H. Huang, Cascaded Feature Network for Semantic Segmentation of RGB-D Images, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, 2017: pp. 1320–1328. doi: 10.1109/ICCV.2017.147.
[35] W. Wang, U. Neumann, Depth-Aware CNN for RGB-D Segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 144–161, https://doi.org/10.1007/978-3-030-01252-6_9.
[36] H. Maheshwari, Y.-C. Liu, Z. Kira, Missing Modality Robustness in Semi-Supervised Multi-Modal Semantic Segmentation, in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2024: pp. 1009–1019. doi: 10.1109/WACV57701.2024.00106.
[37] A. Pemasiri, K. Nguyen, S. Sridharan, C. Fookes, Multi-modal semantic image segmentation, Comput. Vis. Image Underst. 202 (2021) 103085, https://doi.org/10.1016/j.cviu.2020.103085.
[38] T. Zhou, S. Ruan, S. Canu, A review: Deep learning for medical image segmentation using multi-modality fusion, Array 3–4 (2019) 100004, https://doi.org/10.1016/j.array.2019.100004.
[39] B. Lu, J. Zhou, Q. Wang, G. Zou, J. Yang, Fusion-based color and depth image segmentation method for rocks on conveyor belt, Miner. Eng. 199 (2023) 108107, https://doi.org/10.1016/j.mineng.2023.108107.
[40] M. Li, M. Chen, W. Lu, F. Zhao, P. Yan, J. Liu, RDT-FragNet: A DCN-Transformer network for intelligent rock fragment recognition and particle size distribution acquisition, Comput. Geotech. 177 (2025) 106809, https://doi.org/10.1016/j.compgeo.2024.106809.
[41] S.-H. Liu, C.-G. Liu, C.-M. Shen, H.-Y. Mao, J.-C. Zhu, Automated particle size characterization and gradation evaluation of rock fragments for dam construction using close-range photogrammetry, Adv. Eng. Inf. 68 (2025) 103641, https://doi.org/10.1016/j.aei.2025.103641.
[42] H. Fan, Z. Tian, X. Sun, H. Liu, J. Li, J. Xiang, C. Huang, Gradation regression prediction for engineering based on multiscale rockfill instance segmentation, Adv. Eng. Inf. 64 (2025) 103090, https://doi.org/10.1016/j.aei.2024.103090.
[43] Y. Zhang, Y. Ma, Y. Li, L. Wen, Intelligent analysis method of dam material gradation for asphalt-core rock-fill dam based on enhanced Cascade Mask R-CNN and GCNet, Adv. Eng. Inf. 56 (2023) 102001, https://doi.org/10.1016/j.aei.2023.102001.
[44] J. Knodt, Z. Pan, K. Wu, X. Gao, Joint UV Optimization and Texture Baking, ACM Trans. Graph. 43 (2024) 1–20, https://doi.org/10.1145/3617683.
[45] Y. Tang, Y. Wang, G. Si, Vision-based size distribution analysis of rock fragments using multi-modal deep learning and interactive annotation, Autom. Constr. 159 (2024) 105276, https://doi.org/10.1016/j.autcon.2024.105276.
[46] R.D. Moreira, F. Coutinho, L. Chaimowicz, Analysis and Compilation of Normal Map Generation Techniques for Pixel Art, in: 2022 21st Brazilian Symposium on Computer Games and Digital Entertainment (SBGames), 2022, pp. 1–6, https://doi.org/10.1109/SBGAMES56371.2022.9961116.
[47] Y. Tang, Y. Wang, X. Wang, J. Oh, G. Si, Automated Scene-Adaptive Rock Fragment Recognition Based on the Enhanced Segment Anything Model and Fine-Tuning RTMDet, Rock Mech. Rock Eng. 58 (2025) 3973–3999, https://doi.org/10.1007/s00603-024-04360-0.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, (2015). doi: 10.48550/arXiv.1512.03385.
[49] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal, Networks (2016), https://doi.org/10.48550/arXiv.1506.01497.
[50] J. Long, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, (2015). doi: 10.48550/arXiv.1411.4038.
[51] K. He, X. Zhang, S. Ren, J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, in: Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science. pp. 346–361. doi: 10.1007/978-3-319-10578-9_23.
[52] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, M.-M. Cheng, Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks, IEEE Trans. Neural Netw. Learning Syst. 32 (2021) 2075–2089, https://doi.org/10.1109/TNNLS.2020.2996406.
[53] Z. Chen, R. Cong, Q. Xu, Q. Huang, DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection, IEEE Trans. on Image Process. 30 (2021) 7012–7024, https://doi.org/10.1109/TIP.2020.3028289.
[54] T. Vu, H. Kang, C.D. Yoo, SCNet: Training Inference Sample Consistency for Instance Segmentation, AAAI 35 (2021) 2701–2709, https://doi.org/10.1609/aaai.v35i3.16374.
[55] K. Chen, W. Ouyang, C.C. Loy, D. Lin, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, Hybrid Task Cascade for Instance Segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019: pp. 4969–4978. doi: 10.1109/CVPR.2019.00511.
[56] Z. Cai, N. Vasconcelos, Cascade R-CNN: High Quality Object Detection and Instance Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 1483–1498, https://doi.org/10.1109/TPAMI.2019.2956516.
[57] X. Ou, W. Zhou, T. Qu, J. Yang, Uncertainty-informed large vision model for automated recognition of excavated rock chips in tunnel boring machines, J. Rock Mech. Geotech. Eng. (2025), https://doi.org/10.1016/j.jrmge.2025.08.030.