The Visual Computer (2025) 41:11311–11325
https://doi.org/10.1007/s00371-025-04108-z

REVIEW

Advancing landslide recognition through multi-dimensional feature fusion and transformer architectures

Cong Chen[1] · Chengwei Yu[2] · Shanshan Cai[3]

Accepted: 3 July 2025 / Published online: 6 August 2025

Abstract
Landslides are a type of sudden and highly destructive geological hazard. Traditional detection methods often suffer from delayed response and low efficiency. In recent years, deep learning-based object detection techniques have attracted increasing attention in disaster recognition tasks, particularly transformer-based detection models, which exhibit significant advantages in global feature modeling. However, landslide targets in remote sensing imagery often present challenges such as large-scale variation, blurred boundaries, and texture interference. To address these issues, this study proposes an improved detection algorithm based on the RT-DETR-r18 framework by integrating multiple specialized modules. First, the DDC3 module is designed to enhance the recognition of fine boundaries and local textures, thereby improving the feature extraction capacity of the backbone network. Second, an Efficient Additive Attention (EAA) mechanism is introduced to suppress redundant information and strengthen the model's focus on critical regions, improving detection precision. Finally, the CGAFusion module is employed, which utilizes a triple-attention strategy to collaboratively regulate feature weights. This module enhances the model's ability to filter salient features while preserving global contextual information, leading to more accurate landslide edge detection. A dual-class dataset comprising landslides and storms is constructed from multi-source imagery for evaluation. The experimental results show that the proposed method outperforms existing models in several dimensions including mAP@0.5 and F1 score, demonstrating strong detection accuracy. Code is available at https://github.com/sanyauChenCoder/Landslide_02.git.

Keywords RT-DETR-r18 · Landslide detection · DDC3 module · Efficient Additive Attention · CGAFusion module

1 Introduction

Landslides, as a common type of geological disaster, pose serious threats to human life, property, and ecological stability [1]. Under triggering conditions such as heavy rainfall and earthquakes, landslides are often characterized by sudden onset, concealment, and destructive impact. Timely and accurate identification of landslide-affected areas is crucial for disaster early warning and emergency response. Conventional landslide detection methods mainly rely on field surveys or manual interpretation of high-resolution remote sensing imagery. These approaches are typically labor-intensive, time-consuming, and lack automation, making them insufficient for practical applications that demand rapid response and large-scale coverage.

In recent years, with the advancement of deep learning technologies, convolutional neural network (CNN)-based object detection methods have demonstrated significant

Cong Chen and Chengwei Yu are co-first authors.

✉ Shanshan Cai
s.cai6@lancaster.ac.uk

Cong Chen
chencongsanya@126.com

Chengwei Yu
chengweiyu@buaa.edu.cn

[1] School of Marine Information Engineering, Hainan Tropical Ocean University, Sanya 572022, China

[2] China Fire and Rescue Institute, 4 Nanyan Road, Changping District, Beijing 102202, China

[3] Division of Biomedical and Life Sciences, Faculty of Health and Medicine, Lancaster University, Lancaster LA1 4YG, UK

potential in natural disaster monitoring. In particular, one-stage detectors such as the YOLO (You Only Look Once) series have been widely adopted in the recognition of landslides, debris flows, floods, and other disaster scenarios, due to their high detection speed and deployment flexibility. While these methods have achieved notable improvement landslide detection efficiency in practical applications, they still face challenges in complex environments involving structural diversity, large-scale variations, and blurred imagery. These issues often result in missed detections and false positives, revealing limitations in feature representation capabilities.

To tackle the aforementioned challenges, researchers have explored the integration of Transformer architectures to improve global context modeling in detection tasks. Transformers offer inherent advantages in capturing long-range dependencies and contextual relationships, making them well suited for complex scene understanding. RT-DETR (Real-Time Detection Transformer), which integrates Transformer mechanisms with a convolutional backbone, has emerged as a promising framework balancing real-time performance and semantic modeling. For example, Fan et al. [2] introduced the ETGC2-Net architecture, which combines an Enhanced Transformer (EFormer) with a Graph Convolutional Network (GCN), guided by superpixel segmentation to construct graph structures, thereby improving the accuracy and stability of landslide detection. However, the complex structure of the model incurs high deployment costs, limiting its application on resource-constrained platforms. Ren et al. [3] proposed a shallow landslide recognition method based on an improved Otsu algorithm and multi-feature thresholding, which yielded clear results but was restricted by the use of samples from a single region, thus limiting generalization. Chen et al. [4] developed a dual-branch convolution transformer network (CTDNet), which achieved excellent performance on the Landslide4Sense dataset. Nevertheless, its performance is highly dependent on dataset-specific features, raising concerns about cross-region adaptability. Tang et al. [5] employed the SegFormer architecture for earthquake-induced landslide identification, which achieved high detection accuracy, although the computational burden introduced by the Transformer architecture affected real-time applicability. Gao et al. [6] proposed the OMV-HDL method, which integrates YOLOv5 and DETR models through optimal and multi-view fusion strategies, substantially improving the detection of old landslides. However, the method also significantly increased the complexity of training and deployment. Li et al. [7–9] explored transformer-based multi-scale fusion and dictionary-driven unfolding models to improve detail preservation in pan-sharpened imagery, although these methods often incur higher computational cost or exhibit weaker spectral retention. Despite these advances, many existing models suffer from increased complexity, poor adaptability across regions, or insufficient edge detail extraction. Building upon these insights, this study aims to address three persistent challenges in landslide detection: (1) insufficient capture of fine-grained edge features, (2) ineffective suppression of irrelevant background noise, and (3) instability in multi-scale feature representation. To this end, we propose an improved RT-DETR-r18-based architecture incorporating three complementary modules: DDC3 for enhancing spatial detail extraction, EAA for efficient global attention, and CGAFusion for robust multi-scale integration. The main contributions are as follows:

1. First, the DDC3 module is designed to enhance the backbone network's ability to perceive and recognize fine boundaries and local textures in the image, thereby improving the accuracy and completeness of feature representation from the source.
2. Second, the Efficient Additive Attention (EAA) mechanism is introduced to enhance the response strength of key regions through linear complexity attention operations. This effectively suppresses interference from redundant targets, improving detection accuracy and robustness.
3. Finally, the CGAFusion module is introduced, integrating channel attention, spatial attention, and pixel attention mechanisms to achieve collaborative regulation of multi-dimensional information. While ensuring contextual integrity, it improves the selectivity and discriminative power of the fused features, significantly enhancing landslide edge extraction and region segmentation performance.

2 Related work

2.1 Evolution of object detection algorithms: from YOLO to transformer

Object detection technology, as the core engine of landslide recognition models, has evolved from traditional two-stage methods based on sliding windows and manual feature extraction to lightweight and efficient end-to-end detection frameworks. With the rapid development of deep learning, the YOLO (You Only Look Once) series [10–20] has become one of the mainstream solutions for geohazard detection, especially landslide recognition tasks, due to its end-to-end training, high-speed inference, and flexible deployment. YOLOv5 introduced the Cross-Stage Partial (CSP) module in its backbone network to enhance feature representation and improve gradient flow efficiency across the network. YOLOv8 further optimized the backbone and feature fusion networks (Neck), improving detection performance for small objects. These improvements have made the YOLO series models outstanding in most natural image tasks, offering high detection speed and accuracy. However, in high-resolution remote sensing images with sparse targets and complex backgrounds, YOLO models face new challenges. Landslide targets typically exhibit characteristics such as blurred boundaries, irregular shapes, and weak textures, making it difficult for local convolutional receptive fields to effectively model long-range dependencies. This results in missed detections and false positives in practical applications, affecting the overall recognition performance. To overcome the limitations of CNN structures in modeling global relationships, Transformer architectures have gradually been introduced into the field of object detection. DETR (Detection Transformer) was the first to apply the encoder–decoder structure from natural language processing to visual tasks, discarding traditional anchor box mechanisms and leveraging self-attention mechanisms to enhance global modeling capabilities. Subsequent methods, such as Deformable DETR and DN-DETR, optimized training convergence speed and spatial feature localization ability while retaining the advantages of global perception, thus expanding their applicability in real-world tasks.

RT-DETR [21], as an improved version designed for real-time detection scenarios, integrates lightweight convolutional modules and a streamlined Transformer architecture, achieving a good balance between detection speed and modeling capability. This model has shown excellent performance in multi-class natural image recognition tasks and provides a feasible structural foundation for landslide detection in remote sensing fields. However, in practical applications, RT-DETR still faces issues such as insufficient shallow edge feature extraction and insensitivity to blurry targets. These problems are particularly evident when dealing with landslide regions characterized by complex structures and significant scale variations, where recognition performance exhibits fluctuations. Therefore, the key challenge in current research is how to further optimize the structure to enhance the model's ability to finely perceive landslide targets while maintaining its real-time advantages.

2.2 Bidirectional feature pyramid network

BiFPN (Bidirectional Feature Pyramid Network) [22] is a structurally optimized and highly efficient feature pyramid network. The core idea of BiFPN is to establish bidirectional fusion paths between multi-scale features, both top down and bottom up, while introducing a learnable weight mechanism to improve the quality of feature interactions. The network takes feature layers with different resolutions (from P3 to P7) as input, where low-level features contain rich spatial details and high-level features provide stronger semantic abstraction.

In the top-down path, high-level semantic features are passed down to lower levels and fused with low-level features to enhance their semantic representation. In the bottom-up path, low-level features are upsampled and recombined with high-level features to supplement spatial details. In each fusion step, BiFPN uses a learnable weighted addition method instead of the traditional direct summation, ensuring the model can dynamically adjust the importance of features at different scales. Furthermore, this structure strengthens cross-layer information flow through well-designed skip connections, while avoiding redundant computation paths, thus effectively controlling computational costs while improving accuracy.

Through this mechanism, BiFPN builds a more discriminative and adaptable feature pyramid that effectively supports multi-scale object detection tasks. It has demonstrated excellent performance in tasks such as classification and bounding box regression and is considered one of the representative structures in the field of multi-scale feature fusion.

2.3 RT-DETR-r18 model

The core architecture of RT-DETR-r18 consists of four main components: the backbone feature extraction network, hybrid encoder, IoU-aware query selection, and decoder. The backbone network, typically implemented using convolutional neural networks such as ResNet18, is responsible for extracting multi-scale features from the input image. One of the key innovations of RT-DETR-r18 lies in its hybrid encoder, which significantly reduces computational complexity by decoupling intra-scale interactions and cross-scale feature fusion. Specifically, the Attention-based Intra-scale Feature Interaction (AIFI) module within the hybrid encoder applies a single layer of multi-head self-attention only to the last stage of features output by the backbone, in order to capture high-level semantic information. Meanwhile, the Cross-scale Complementary Feature Merging (CCFM) module adopts a structure similar to FPN + PAN, fusing the output of the AIFI module with the second-to-last and third-to-last feature maps from the backbone to produce enhanced multi-scale feature representations.

3 Method design

This study proposes an enhanced landslide detection algorithm based on the RT-DETR-r18 framework to improve the detection accuracy of landslide targets in complex natural scenes. Instead of simple modular stacking, the three components—DDC3, EAA, and CGAFusion—are integrated with complementary functions following a top-down structural

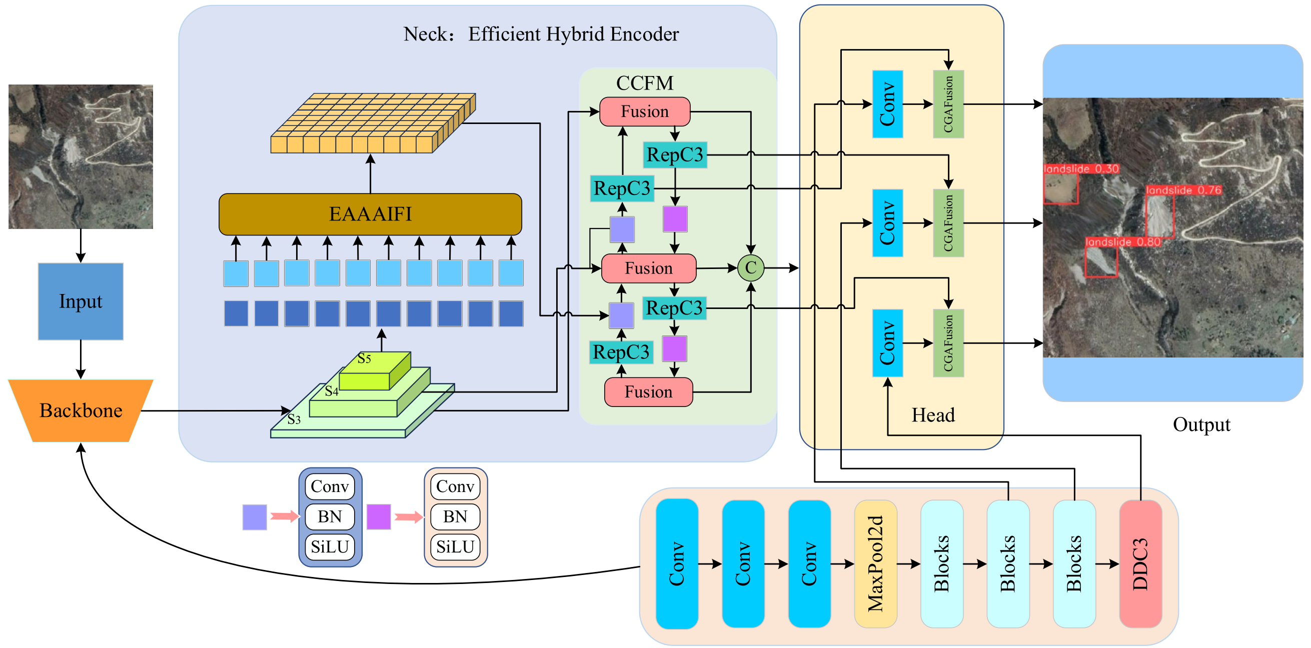


Fig. 1 Improved RT-DETR-r18 network architecture

logic. These modules are designed to enhance low-level edge and texture features, suppress irrelevant noise through global attention mechanisms, and refine multi-scale feature representations via collaborative attention. The integration maintains the real-time inference capability of RT-DETR-r18 while systematically optimizing three key aspects: backbone feature extraction, global semantic modeling, and multi-scale feature fusion. This design accommodates the variation in scale, morphology, and semantic complexity characteristic of landslide targets. The DDC3 module utilizes depthwise separable convolutions and cross-channel residual connections to enhance the extraction of local texture and edge information in the backbone. The EAA module introduces an additive attention mechanism that enables efficient global context modeling with reduced computational complexity, addressing the challenge of long-range dependencies and blurred boundaries in landslide imagery. The CGAFusion module incorporates spatial, channel, and pixel-level attention mechanisms to alleviate feature misalignment and semantic confusion during multi-scale fusion, thereby improving the discrimination of target regions. Figure 1 illustrates the overall architecture of the improved RT-DETR-r18 network. The network processes the input image through the DDC3-enhanced backbone, applies global semantic refinement via the EAA module within the Transformer encoder–decoder structure, and performs final multi-scale integration through the CGAFusion module to generate detection and segmentation outputs. The design emphasizes a sequential, hierarchical flow of features and establishes functional coordination among the modules to improve overall detection performance.

3.1 DDC3 module

In landslide-related remote sensing recognition tasks, targets often exhibit characteristics such as blurred boundaries and complex texture variations, which limit the ability of traditional convolutional neural networks to extract multi-scale contextual information and local structural details. To enhance the feature representation capability of the model, this study proposes the DDC3 module, designed to improve the model's ability to recognize geophysical targets while maintaining network lightweight.

As shown in Fig. 2, the input feature X first undergoes a standard convolutional layer for preliminary transformation, adjusting the channel count and extracting low-level local features. The feature then passes through the main path, composed of n cascaded DWR-DRB [23, 24] modules. Each DWR-DRB module adaptively enhances contextual information at different receptive fields while boosting local feature extraction capabilities. After passing through multiple DWR-DRB modules, the main branch integrates the extracted features through a convolutional layer. Simultaneously, the input also passes through a shortcut branch, using a single convolution to match the channel dimensions. The outputs of the main and shortcut branches are fused by element-wise addition, and the fused features undergo a 1 ×

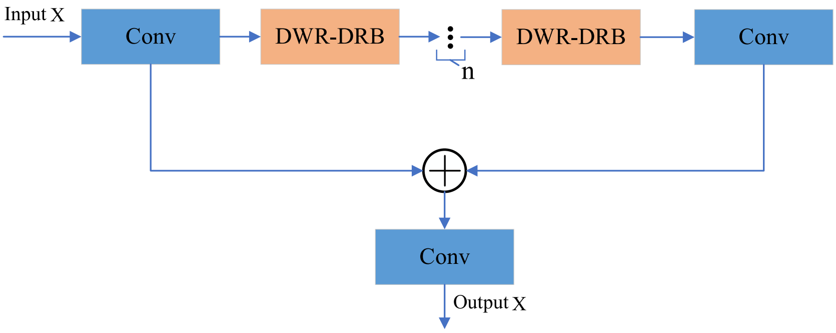


Fig. 2 DDC3 module

1 convolution for channel projection, resulting in the final output feature X.

3.2 Efficient additive attention module

In landslide remote sensing images, the target–background interference is complex, which places high demands on the model's global modeling and key region perception capabilities. Although traditional self-attention mechanisms possess global perception abilities, they incur high computational costs. To improve modeling efficiency and feature focusing capability, this study introduces the Efficient Additive Attention module. By replacing complex multiplication operations with an additive attention mechanism, the module utilizes learnable vectors for efficient global feature extraction, enhancing the model's response to target and edge regions while reducing computational overhead.

Efficient Additive Attention [25] is an efficient attention mechanism that reduces computational complexity from quadratic to linear by avoiding expensive matrix multiplication operations, making it particularly suitable for real-time visual applications on resource-constrained mobile devices. As shown in Fig. 3, its implementation process is as follows: First, the input feature matrix x is transformed into query and key matrices Q and K, implemented through matrices W_q and W_k, with parameters Q, $K \in \mathbb{R}^{n\times d}$, W_q, and $W_k \in \mathbb{R}^{d\times d}$. Then, the query matrix Q is multiplied by a learnable parameter vector w_a, learning the attention weights of the query to generate the global attention query vector $\alpha = Q \cdot w_a/\sqrt{d}$. Next, based on the learned attention weight α, the query matrix Q undergoes weighted pooling to obtain a single global query vector $q = \sum_{i=1}^{n} \alpha_i Q_i$. Afterward, element-wise multiplication is performed to allow interaction between the global query vector q and the key matrix K, forming the global context representation $\hat{x} = Q + T(K * q)$, where T is a linear transformation. The final output representation is the sum of the query matrix Q and the global context representation learned through the linear transformation.

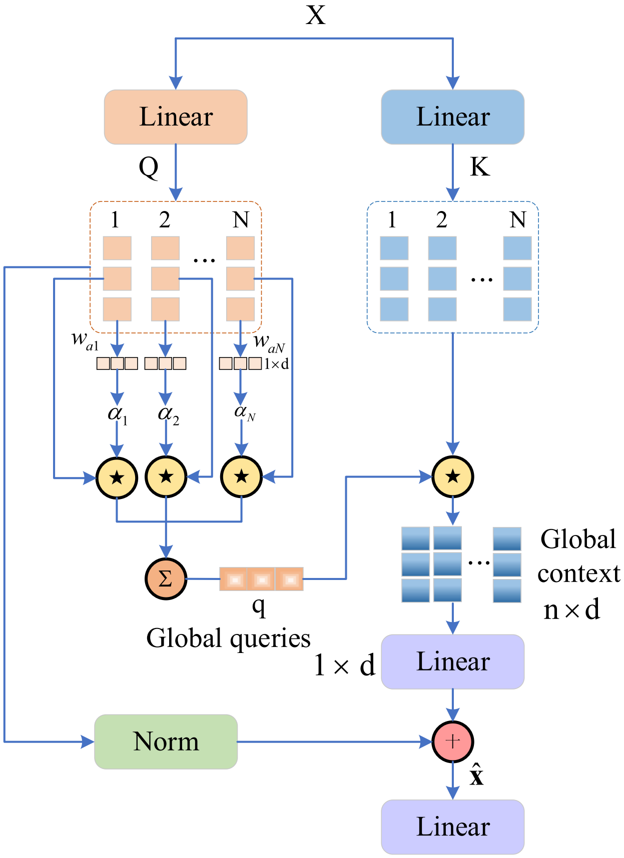


Fig. 3 Efficient Additive Attention module

3.3 CGAFusion module

To further enhance the information interaction efficiency and fusion quality between multi-source features, this study introduces the CGAFusion module. This module employs spatial attention (SA), channel attention (CA), and pixel attention (PA) mechanisms to achieve guided fusion across branches, effectively strengthening the collaborative expression ability between shallow details and deep semantics.

As shown in Fig. 4, the implementation process of the CGAFusion [26] module includes the following steps: First, the CGA module generates spatial importance maps (SIMs)

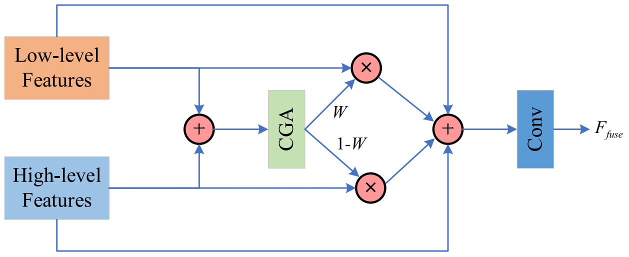


Fig. 4 CGAFusion module

for each channel. The CGA module utilizes spatial attention and channel attention mechanisms to generate spatial attention weights and channel attention weights, which are then fused to obtain a rough SIM. Next, guided by the content of the input features, operations such as channel shuffling and group convolutions are applied to refine the rough SIM, generating channel-specific fine SIMs. Subsequently, both low-level and high-level features from the encoder are input into the CGA module to obtain spatial weights, and these weights are used to compute a weighted sum of the low-level and high-level features. Then, a residual connection is used to add the input features to the fused features, enhancing feature information flow, mitigating the gradient vanishing problem, and simplifying the learning process. Finally, the fused features are projected through a 1 × 1 convolutional layer to obtain the final features. Through this process, CGA fusion effectively integrates both detail and semantic information within the image.

4 Experiment

4.1 Dataset

The dataset used in this study consists of two main categories: landslides and storms, with all images manually annotated to ensure accuracy and consistency. Existing landslide recognition datasets often rely on high-resolution remote sensing images; however, such data struggle to realistically reflect the complex environmental features at the time of a landslide event, as shown in the sample images in Fig. 5. Additionally, the processing of remote sensing images requires substantial computational resources, limiting their feasibility in real-time applications. To overcome these issues, this study constructed a multi-domain landslide image dataset consisting of a total of 4,735 images. Of these, 77.8% are frame-extracted images from news reports and live videos, while 22.2% are drone remote sensing images, enhancing the model's generalization and adaptability in complex natural scenes. Given the high-speed sliding characteristics of the landslide process, the data construction used a frame extraction strategy with frame extraction intervals set at 1 s, 1.5 s, and 2 s to preserve key dynamic features. Furthermore, considering that some images suffer from issues such as watermarks, blurriness, or occlusions by buildings, 1125 images were subject to offline augmentation, representing 24% of the total data. This dataset offers significant advantages in terms of dynamics, realism, and diversity, providing a more representative set of training samples for landslide detection tasks in complex terrain conditions.

4.2 Experimental platform

The experiments were conducted on a well-equipped computing platform, using the deep learning framework PyTorch 1.10.0 with Python 3.8, running on an Ubuntu 20.04 system environment. GPU acceleration was implemented through CUDA 11.3. During training, a single NVIDIA RTX 4090 GPU with 24GB of memory was used, along with an AMD EPYC 7T83 64-core processor (including 22 virtual cores) and 90 GB of system memory. This setup provided efficient computational power, accelerating the model's convergence speed and overall training process.

4.3 Hyperparameter setting

All experiments were conducted with an input resolution of 640 × 640 pixels to balance detection accuracy and computational cost. The model was trained for 200 epochs using a batch size of 16. Stochastic gradient descent (SGD) was adopted as the optimizer, with an initial learning rate of 0.01, momentum set to 0.937, and a weight decay factor of 0.0005. The learning rate followed a cosine annealing schedule with warmup during the first 3 epochs. The number of worker threads for data loading was set to 4 to ensure efficient throughput. These settings were applied consistently across all experiments and baseline comparisons to ensure fairness and reproducibility.

4.4 Evaluation metrics

In this study, the evaluation metrics used include F1 score, precision (P), recall (R), average precision (AP), and mean average precision (mAP) [27]. Additionally, the number of parameters (Parameters) was also considered. The formulas for these metrics are as follows:

$$\text{Precision} = \frac{T_\text{p}}{T_\text{p} + F_\text{p}} \tag{1}$$

$$\text{Recall} = \frac{T_\text{p}}{T_\text{p} + F_\text{N}} \tag{2}$$

$$\text{AP} = \int_0^1 P(R)\text{d}R \tag{3}$$

Fig. 5 Image sample

$$\mathrm{mAP} = \frac{1}{n}\sum_{\mathrm{i}=0}^{n} \mathrm{AP}(i) \tag{4}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5}$$

where T_{p} denotes the number of correctly detected targets, F_{p} represents the number of incorrectly detected targets, F_{N} is the number of missed targets, n is the number of classes, and $\mathrm{AP}(i)$ represents the average precision of the iii-th target class.

5 Experimental analysis

5.1 Algorithm comparison results

To further validate the effectiveness and generalization ability of the proposed model, this study conducts comparative experiments with several representative object detection models. The selected baseline models include multiple versions of the YOLO series (YOLOv5m, YOLOv5l, YOLOv8m, YOLOv8l, YOLOv10l) and the RT-DETR-r18 model based on the Transformer architecture, covering both convolution-based and attention-based detection methods. To ensure a fair comparison, all models are trained and evaluated under the same training parameters and dataset settings. The evaluation metrics include precision, recall, mean average precision at IoU 0.5 (mAP@0.5), and F1 score, which comprehensively measure detection accuracy, robustness, and the balance between false positives and missed detections. The comparative results are summarized in Table 1.

Table 1 shows that our method achieves the highest performance among all compared models, with accuracy, recall, and mAP@0.5 reaching 76.5%, 67.4%, and 69.7%, respectively, and an F1 score of 72.0%. Compared to YOLOv5m and YOLOv5l, our method improves mAP by 1.0% and 0.7%, respectively. Although YOLOv10l has a slightly higher accuracy (77.4%), its recall significantly drops to 52.9%, with an mAP of only 59.8% and an F1 score reduced to 63.0%, resulting in overall poorer detection performance. The performance of the YOLOv8 series models varies across metrics. YOLOv8m shows relatively balanced performance, but its mAP (68.1%) is lower than that of our method. Although YOLOv8l has higher accuracy (75.9%), its recall is lower (59.1%), limiting its overall detection accuracy. RT-DETR-r18 matches our method in recall (67.4%) but has lower accuracy (74.3%) and slightly lower mAP (66.2%), indicating shortcomings in its overall detection accuracy and feature representation capability. The comprehensive comparison results demonstrate that our method, while maintaining high accuracy, effectively balances recall capability, showing stronger detection robustness and generalization across multiple target categories and scales. The leading advantage in the F1 score further confirms the stability and effectiveness of the proposed method in practical detection scenarios.

To further validate the detection capability of the proposed model on different target types, this study evaluates the typical categories "landslide" and "storm" for single-class assessments and compares the detection accuracy of the proposed model with that of mainstream models for each category. The results are shown in Table 2. In the "landslide" category, YOLOv5l (64.2%) and YOLOv5m (63.8%) show similar performance, with YOLOv8m achieving a detection accuracy of 63.1%, slightly lower than the YOLOv5 series. YOLOv8l further decreases to 59.4%, and YOLOv10l performs the weakest in this category, with a score of only 52.4%. RT-DETR-r18 scores 62.0%, placing it at a moderate level. In contrast, the proposed method achieves a detection accuracy of 65.0% in this category, outperforming all the compared models and achieving the best performance. In the "storm" category, the YOLOv8 series overall shows

Table 1 Accuracy comparison of different object detection models

Algorithm	Precision/%	Recall/%	mAP@0.5/%	F1/%
YOLOv5m	75.5	66.3	68.0	71.0
YOLOv5l	72.2	68.5	69.0	70.0
YOLOv8m	71.3	64.8	68.1	68.0
YOLOv8l	75.9	59.1	65.9	67.0
YOLOv10l	77.4	52.9	59.8	63.0
RT-DETR-r18	74.3	67.4	66.2	70.0
Ours	76.5	67.4	69.7	72.0

Table 2 Average precision comparison (AP%) of different models on landslide and storm targets

Classes	Algorithms						
	YOLOv5m	YOLOv5l	YOLOv8m	YOLOv8l	YOLOv10l	RT-DETR-r18	Ours
landslide	63.8	64.2	63.1	59.4	52.4	62.0	65.0
storm	72.2	73.8	73.1	72.4	67.2	70.4	74.5

stable performance, with YOLOv8m and YOLOv8l achieving detection accuracies of 73.1% and 72.4%, respectively. YOLOv5m and YOLOv5l achieve detection accuracies of 72.2% and 73.8%, slightly higher than the YOLOv8 series. YOLOv10l also experiences a significant performance drop in this category, with a score of only 67.2%. RT-DETR-r18 performs at 70.4%. Among all models, the proposed method achieves 74.5% in this category, which is the best result. In summary, the proposed method achieves the highest detection accuracy for both landslide and storm categories, demonstrating excellent detection capability and robustness.

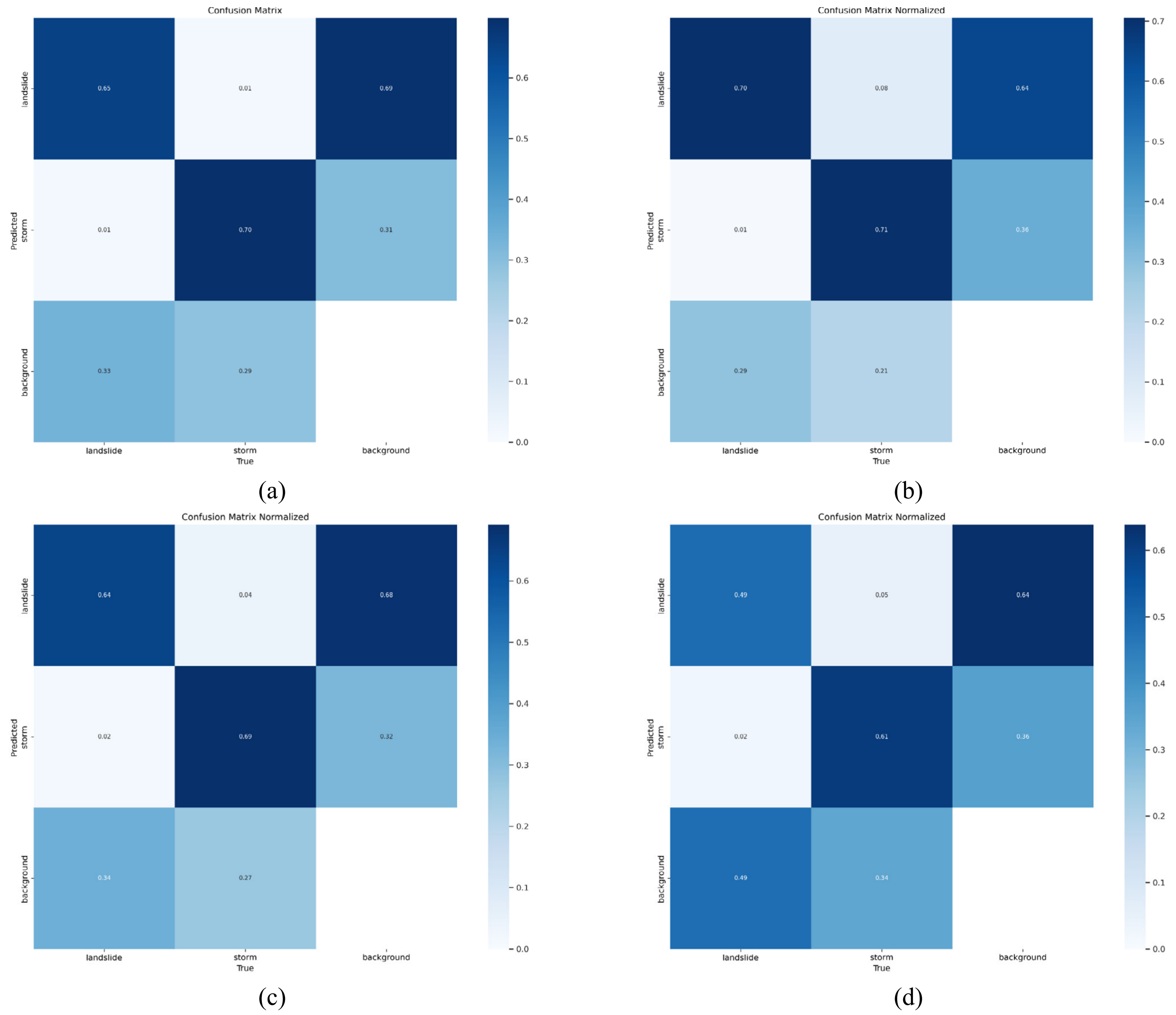
5.2 Result visualization

To further validate the classification performance of each detection model on remote sensing images and their ability to distinguish between different target categories, this study introduces visualization techniques to analyze the model's prediction results. A normalized confusion matrix is used to compare and display the classification performance of YOLOv5l, YOLOv8m, YOLOv8l, YOLOv10l, RT-DETR-r18, and the proposed method on the two core categories: landslide and storm. The confusion matrix effectively reflects the model's accuracy in category recognition and the sources of errors, helping to reveal potential biases or limitations that the model may encounter in practical applications.

To further compare the classification performance and false detection control ability of each model on target categories, this study presents the normalized confusion matrix diagrams, as shown in Fig. 6. The compared models include YOLOv5l, YOLOv8m, YOLOv8l, YOLOv10l, RT-DETR-r18, and Ours, with the analysis focusing on the two core target categories: landslide and storm. In the landslide category recognition, YOLOv10l performs the weakest with an accuracy of only 0.49, misclassifying a large portion of the background area as the target region, indicating low boundary discrimination ability. YOLOv5l, the YOLOv8 series, and RT-DETR-r18 have landslide recognition accuracies in the range of 0.64–0.75, but the false detection rates remain high. In contrast, Ours achieves an accuracy of 0.75 in this category, with background misclassification controlled below 0.24, significantly outperforming other models and demonstrating stronger structural analysis and spatial suppression capabilities. In the storm category, YOLOv10l also faces severe background confusion, with a background misclassification rate of 0.34, whereas the proposed method reduces this value to 0.21, with a storm class recognition accuracy of 0.74, the highest among all models. This shows that the proposed model not only leads in classification accuracy but also exhibits a significant advantage in false detection suppression, effectively reducing the interference from background noise, minimizing false detections, and improving detection stability. Overall, Ours achieves the optimal detection performance in both major categories.

To validate the effectiveness of the proposed improved structure in enhancing detection performance, this study compares the RT-DETR-r18 model before and after improvement and visualizes its detection performance using precision–recall (P–R) curves. The P–R curve, as an important metric for evaluating object detection quality, reflects the variation in precision at different recall rates, providing a comprehensive assessment of the impact of the improved structure on the model's performance.

To further validate the improvement in detection performance with the proposed module, this study plots the

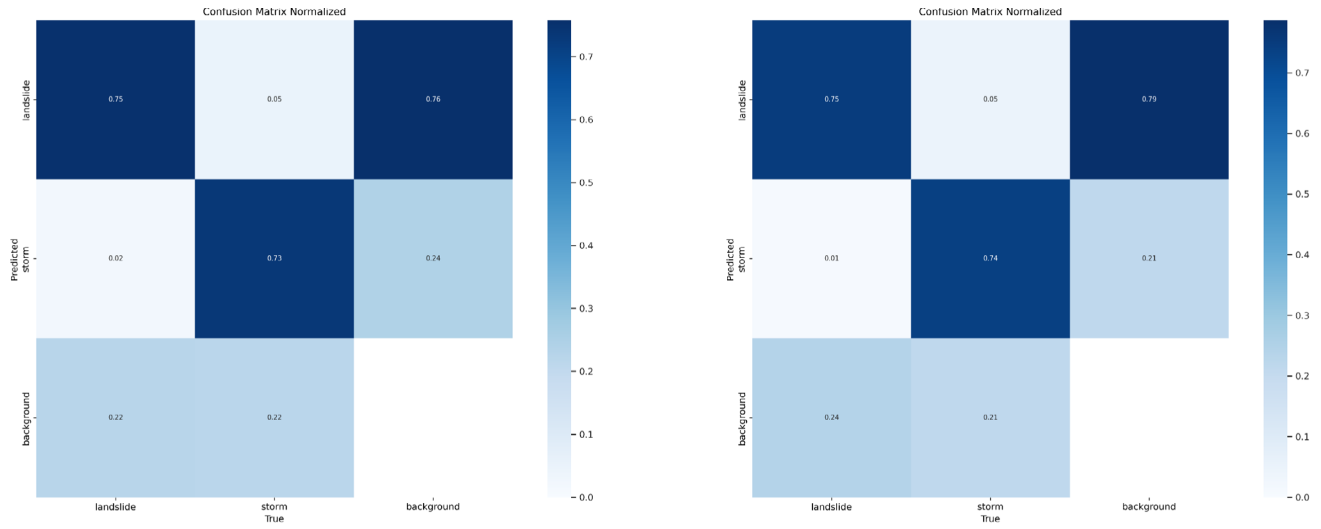


(a) (b) (c) (d)

Fig. 6 Comparison of normalized confusion matrices for each model on landslide and storm categories: **a** YOLOv5l; **b** YOLOv8m; **c** YOLOv8l; **d** YOLOv10l; **e** RT-DETR-r18; **f** Ours

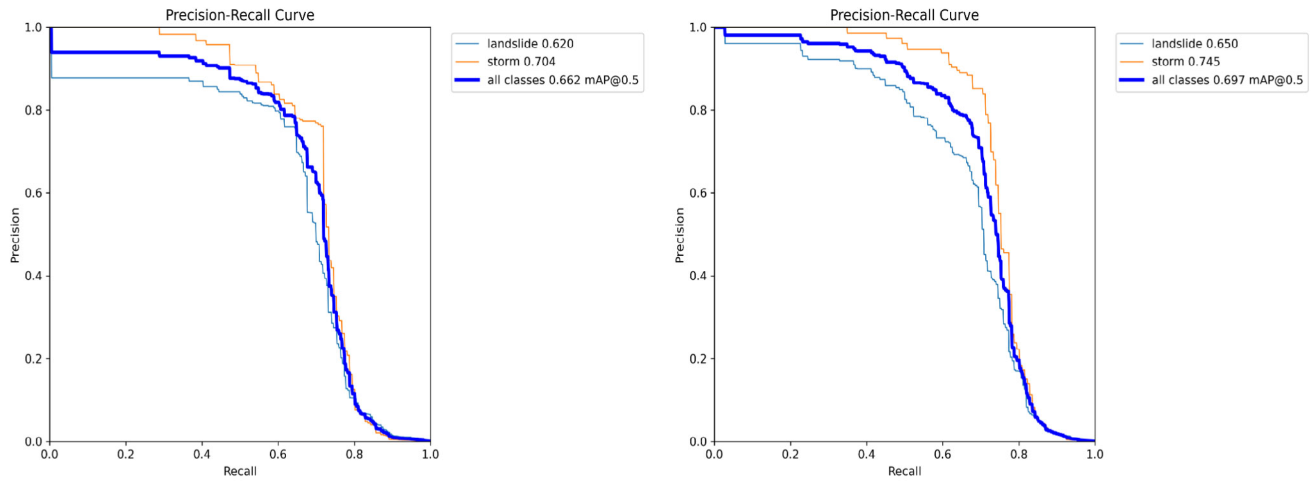
precision–recall curves of the RT-DETR-r18 model before and after enhancement, as shown in Fig. 7. The P–R curve provides a more intuitive reflection of the model's precision variation at different recall rates and quantitatively demonstrates the stability and overall performance of object detection. As seen in the figure, before the improvement, the model's mAP for the landslide and storm categories was 0.620 and 0.704, respectively, with an overall mAP@0.5 of 0.662. After introducing the improved structure, the precision of both categories increased to different extents. Specifically, landslide precision increased to 0.650, storm precision increased to 0.745, and the overall mAP@0.5 rose to 0.697, with an improvement of 3.5%. Notably, in the high recall rate range, the improved model's curve declines more slowly, indicating stronger stability. The results show that the proposed improvement module effectively enhances the model's detection accuracy across different target categories and strengthens its ability to recognize small and boundary targets in complex scenes.

To further validate the recognition capability of different object detection algorithms in real-world landslide scenarios, relying solely on quantitative metrics (such as mAP, Precision, and Recall) has certain limitations. These metrics may fail to comprehensively reveal the model's performance in terms of target boundaries, background interference, and category discrimination in multi-source heterogeneous images. Therefore, this study conducts a comparative analysis of detection results in typical image samples through visualization, aiming to intuitively present the strengths and



(e) (f)

Fig. 6 continued

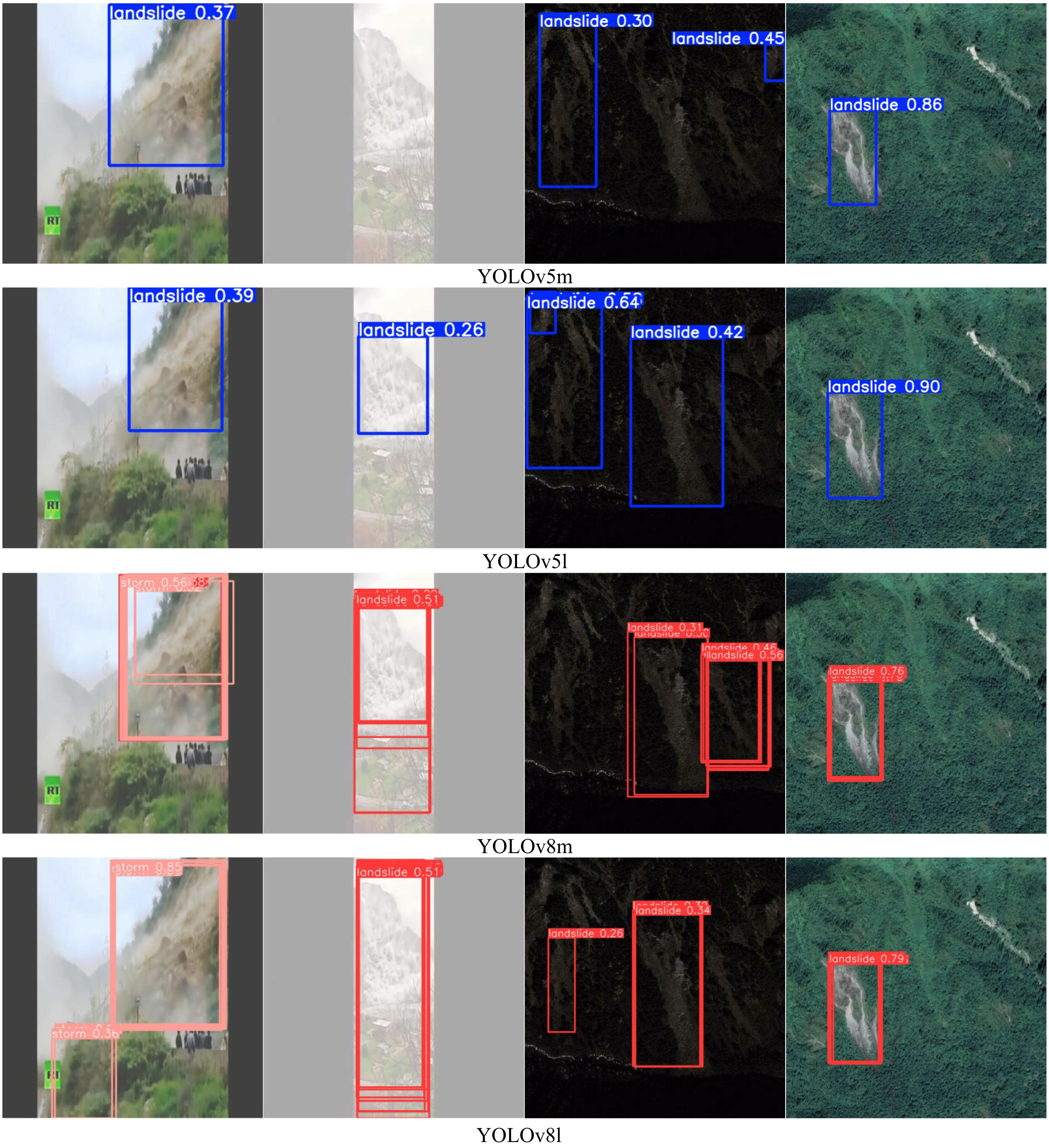


(a) (b)

Fig. 7 Precision–recall curves of RT-DETR-r18 before and after enhancement: **a** RT-DETR-r18; **b** improved RT-DETR-r18

weaknesses of each algorithm in real-world applications from an image-level perspective. The focus is particularly on the model's responsiveness, missed detection issues, and redundant detection problems in low-quality video frames, highly complex remote sensing images, and multi-target mixed scenarios. This approach provides perceptual insights for subsequent model optimization and practical deployment.

As shown in Fig. 8, the detection performance of YOLOv5m, YOLOv5l, YOLOv8m, YOLOv8l, YOLOv10l, RT-DETR-r18, and Ours is presented on multi-source images. In the left two columns, the YOLOv5 series exhibits weak responses to the landslide regions, with generally low confidence scores and missed detections. For instance, in Image 1, YOLOv5m only detects the landslide region with a confidence score of 0.37, while other significant sliding bodies are not captured. Furthermore, YOLOv8m, YOLOv8l, YOLOv10l, and RT-DETR-r18 incorrectly identify the landslide targets as storms, showing significant category confusion. RT-DETR-r18 detects stably in remote sensing images but also exhibits redundant bounding boxes and low confidence repeated annotations in video frames. For example, in Image 2, multiple bounding boxes with confidence scores below 0.3 are generated, causing unnecessary repeated detections that affect practicality. In contrast, Ours demonstrates better robustness and accuracy across multiple scenarios. The model not only accurately detects landslide regions in remote sensing images but also maintains high confidence and fewer redundant bounding boxes



YOLOv5m

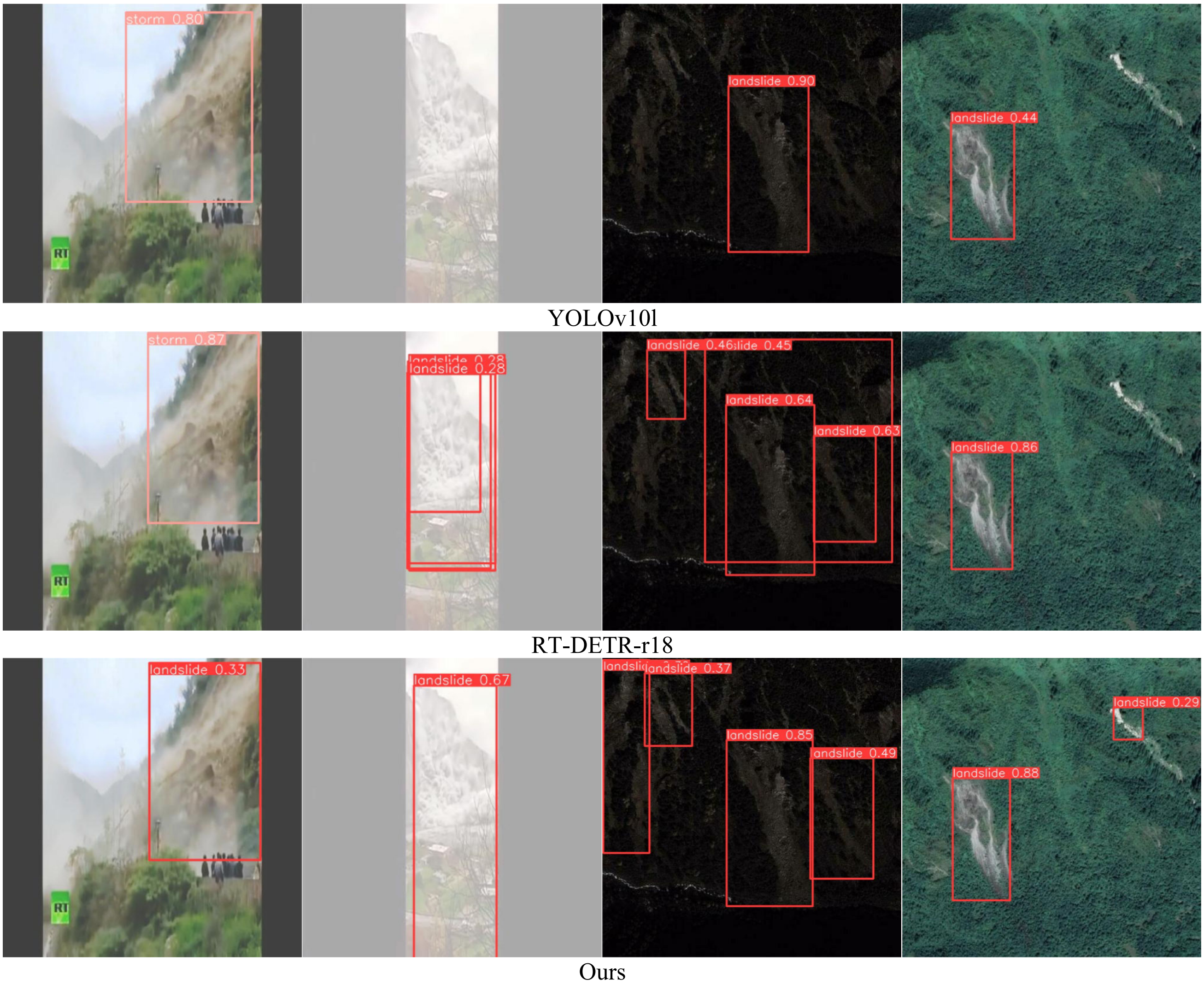
YOLOv5l

YOLOv8m

YOLOv8l

Fig. 8 Detection performance of different algorithms on the landslide dataset

in video images. For instance, in Image 4, the model accurately locates the forest landslide region with a confidence of 0.88 and avoids overlapping bounding boxes or false annotations, outperforming baseline methods such as YOLOv5m, YOLOv8l, and RT-DETR-r18. Additionally, the proposed method exhibits stronger feature representation ability when handling multi-scale targets and complex background interference, especially in multi-target dense or edge-blurred scenarios, showing lower missed detection rates and better box overlap accuracy. Some detection algorithms are prone to misjudgments, false alarms, or redundant boxes in environments with occlusion, smoke interference, and other factors,



YOLOv10l

RT-DETR-r18

Ours

Fig. 8 continued

while our model effectively alleviates these issues. In summary, the visualization results demonstrate that the proposed method outperforms existing mainstream methods in terms of accuracy, localization precision, and detection stability, with stronger adaptability to various scenarios, providing more reliable technical support for intelligent landslide target recognition.

5.3 Ablation study

To further validate the specific contribution of each module to the overall detection performance, this study designs four ablation experiments, progressively introducing the three core modules: the CGAFusion multi-dimensional collaborative attention module, the DDC3 backbone enhancement module, and the EAA efficient additive attention module. Under the condition of keeping other network structures and training parameters consistent, each improvement is evaluated for its effect on the model's accuracy enhancement in landslide and storm target detection.

Table 3 presents the results of the ablation experiments for the four model configurations. Experiment 1 represents the baseline model, RT-DETR-r18, without any improvement modules, serving as the performance comparison baseline. In Experiment 2, the CGAFusion module is added to the baseline model, enabling multi-dimensional attention capabilities in the feature fusion stage, including channel, spatial, and pixel attention. The detection accuracy improves from 66.2% to 69.2%, with a significant increase of 3.0%, indicating that the CGAFusion module plays a key role in cross-scale semantic alignment and edge detail recognition. In Experiment 3, the DDC3 module is further introduced to enhance the backbone network's spatial feature extraction capabilities. Although accuracy slightly decreases from 80.2 to

Table 3 Ablation experiment results

Number	Experiment	Precision/%	Recall/%	mAP@0.5/%
1	RT-DETR-r18	74.3	67.4	66.2
2	RT-DETR-r18 + CGAFusion	80.2	66.6	69.2
3	RT-DETR-r18 + CGAFusion + DDC3	75.2	68.4	69.3
4	RT-DETR-r18 + CGAFusion + DDC3 + EAA	76.5	67.4	69.7

75.2%, the recall rate improves to 68.4%, and mAP continues to rise to 69.3%. This suggests that the DDC3 module improves the perception integrity of landslide regions and offers better adaptability to targets with blurred boundaries. Experiment 4 introduces the EAA module on top of the previous improvements to further strengthen the global semantic modeling ability. Without significantly increasing computational cost, the EAA module effectively enhances the model's understanding of scene context relationships, leading to a final mAP of 69.7%, with precision and recall of 76.5% and 67.4%, respectively, achieving the best overall performance across the four experiments. These results validate the synergistic benefits of the three proposed modules in landslide detection tasks. CGAFusion focuses on local feature fusion, DDC3 strengthens spatial structure modeling, and EAA enhances global context representation. Together, these modules improve the model's detection accuracy and robustness. The progressive improvements observed across experiments suggest that the modules do not function in isolation but build upon each other's outputs. The DDC3 module provides enhanced foundational features, which are more effectively leveraged by the EAA's selective global focus mechanism. The CGAFusion module, in turn, synthesizes these enriched and focused features into a cohesive representation. This layered synergy contributes to both performance robustness and generalization capability, reflecting a deliberate architectural design rather than arbitrary module stacking.

Table 4 presents the performance variations of the RT-DETR-r18 model as the CGAFusion, DDC3, and EAA modules are gradually integrated. The original model exhibits the highest frame rate, the lowest computational load, and the smallest number of parameters, serving as a lightweight and high-speed baseline. After introducing CGAFusion, the model's representational capacity is enhanced, but the increased structural complexity leads to a decrease in frame rate to 161.2 FPS, along with slight increases in GFLOPS and parameters. With the addition of the DDC3 module, GFLOPS increases to 60.1 and parameters grow to 21.6 million, yet the frame rate remains unchanged, indicating that the model maintains speed while gaining stronger feature representation. Finally, the integration of the EAA module results in a slight recovery of frame rate to 166.6 FPS, with a marginal increase in GFLOPS to 60.3 and no further growth in parameters, suggesting the module improves efficiency or feature utilization. Overall, although model complexity increases step by step, inference speed is well preserved. The final version achieves a balance between improved accuracy potential and real-time performance, making it suitable for applications requiring both speed and effectiveness.

6 Conclusion

In response to the high-risk and highly sudden nature of landslides, a geological disaster, this study proposes a landslide target detection algorithm based on an improved RT-DETR-r18 structure. To address issues such as blurred target boundaries, significant scale variations, and complex background interference in landslide images, key modules were designed from the perspective of network structure optimization to enhance detection performance and model robustness. In the backbone network, a DDC3 module is designed to effectively enhance the model's ability to extract fine-grained features. For global perception, the Efficient Additive Attention (EAA) module is proposed to achieve global context modeling at a lower computational cost, improving the representation accuracy of landslide regions. In the feature fusion stage, the CGAFusion module is introduced, integrating spatial, channel, and pixel-level attention to effectively enhance the model's feature response capability to critical regions, improving both global semantics and landslide boundary localization and region segmentation accuracy. The experimental section constructs a multi-source landslide dataset based on real remote sensing imagery and video frame data. Through comparisons with various mainstream detection models such as the YOLO series and RT-DETR, the proposed method demonstrates leading performance in mAP@0.5, F1 score, and false detection control. Further visualization and confusion matrix analysis show that the method exhibits stronger discriminative ability and stability in recognizing both landslide and storm targets. Future research could further explore: (1) more lightweight attention mechanisms to adapt to edge devices, (2) time-series modeling for early landslide warning, and (3) multi-task joint detection strategies for landslides and multiple types of disaster targets.

Table 4 Analysis of the impact of different module integration on the computational efficiency and parameter quantity of the model

number	algorithms	FPS (f/s)	GFLOPS	Params
1	RT-DETR-r18	181.8	56.9	19,874,328
2	RT-DETR-r18 + CGAFusion	161.2	59.2	20,428,193
3	RT-DETR-r18 + CGAFusion + DDC3	161.2	60.1	21,603,233
4	RT-DETR-r18 + CGAFusion + DDC3 + EAA	166.6	60.3	21,603,489

This study provides technical support for intelligent perception and emergency response to geological disasters and has promising practical application prospects.

Author contributions Cong Chen was responsible for the overall framework design, optimization of the deep learning model, and experimental analysis, as well as drafting the initial manuscript. Chengwei Yu contributed to the preparation of the dataset, model implementation and tuning, and performed in-depth analysis of the experimental results, providing key technical support. Shanshan Cai was involved in the design and optimization of certain algorithm modules, assisted in setting up experiments, contributed to the proofreading and revisions of the manuscript, and is the corresponding author for the paper.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Casagli, N., Intrieri, E., Tofani, V., et al.: Landslide detection, monitoring and prediction with remote-sensing techniques. Nat. Rev. Earth Environ. **4**(1), 51–64 (2023)
2. Fan, S., et al.: ETGC2-net: an enhanced transformer and graph convolution combined network for landslide detection. Nat. Hazards **121**(1), 135–160 (2025)
3. Ren, J., et al.: Remote sensing identification of shallow landslide based on improved otsu algorithm and multi feature threshold. Front. Earth Sci. **12**, 1473904 (2024)
4. Chen, X., et al.: Conv-trans dual network for landslide detection of multi-channel optical remote sensing images. Front. Earth Sci. **11**, 1182145 (2023)
5. Tang, X., et al.: Automatic detection of coseismic landslides using a new transformer method. Remote Sens. **14**(12), 2884 (2022)
6. Gao, S., Xi, J., Li, Z., et al.: Optimal and multi-view strategic hybrid deep learning for old landslide detection in the loess plateau, Northwest China. Remote Sens. **16**(8), 1362 (2024)
7. Li, Z., Li, J., Ren, L., et al.: Transformer-based dual-branch multiscale fusion network for pan-sharpening remote sensing images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **17**, 614–632 (2023)
8. Li, Z., Yuan, G., Li, J.: DUCD: deep unfolding convolutional-dictionary network for pansharpening remote sensing image. Expert Syst. Appl. **249**, 123589 (2024)
9. Li, Z., Gao, Y., Yuan, G., et al.: CDME: Convolutional Dictionary Itrative Model For Pansharpening with Mixture-of-Experts. IEEE Geosci. Remote Sens. Lett. (2025). https://doi.org/10.1109/LGRS.2025.3545472)
10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection.
11. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, https://doi.org/10.1109/cvpr.2017.690
12. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. Apr. 08, 2018, arXiv: arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767
13. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection. Apr. 23, 2020, arXiv: arXiv:2004.10934. https://doi.org/10.48550/arXiv.2004.10934
14. Wang, T., Zhai, Y., Li, Y., et al.: Insulator defect detection based on ML-YOLOv5 algorithm. Sensors **24**(1), 204 (2023)
15. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 7464–7475 (2023)
16. Sohan, M., Ram, T.S., Rami Reddy, C.V.: A review on YOLOv8 and Its advancements. Algorithms Intell Syst (2024). https://doi.org/10.1007/978-981-99-7962-2_39
17. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: YOLOv9: learning what you want to learn using programmable gradient information. Feb. 29, 2024. arXiv: arXiv:2402.13616. https://doi.org/10.48550/arXiv.2402.13616
18. Wang, A., et al.: YOLOv10: Real-time end-to-end object detection. Oct. 30, 2024, arXiv: arXiv:2405.14458.
19. Khanam, R., Hussain, M.: YOLOv11: an overview of the key architectural enhancements. Oct. 23, (2024). arXiv: arXiv:2410.17725.
20. Tian, Y., Ye, Q., Doermann, D.: Yolov12: Attention-centric real-time object detectors (2025). arXiv preprint arXiv:2502.12524.
21. He, W., et al.: Object detection for medical image analysis: insights from the RT-DETR model. arXiv preprint arXiv:2501.16469 (2025).
22. Li, N., et al.: Enhanced YOLOv8 with BiFPN-SimAM for precise defect detection in miniature capacitors. Appl. Sci. **14**(1), 429 (2024)
23. Wei, H., et al.: DWRSeg: rethinking efficient acquisition of multi-scale contextual information for real-time semantic segmentation. arXiv preprint arXiv:2212.01173 (2022)

24. Ding, X., et al.: UniRepLKNet: a universal perception Large-Kernel ConvNet for audio video point cloud time-series and image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
25. Shaker, A., et al.: Swiftformer: efficient additive attention for transformer-based real-time mobile vision applications. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
26. Gu, Y., et al.: A conditionally parameterized feature fusion U-net for building change detection. Sustainability **16**(21), 9232 (2024)
27. Kou, R., Wang, C., Peng, Z., et al.: Infrared small target segmentation networks: a survey. Pattern Recogn. **143**, 109788 (2023)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cong Chen is currently pursuing a master's degree at the school of marine information engineering, Hainan Tropical Ocean University, in Sanya, China. His research interests include remote sensing image process and object detection.

Chengwei Yu is lecturer at China Fire and Rescue Academy, graduated with a bachelor's degree from Beihang University in 2018 and a doctoral degree from Beihang University in 2023. His research focuses on PDEs, complex networks, artificial intelligence, and higher education.

Shanshan Cai is a PhD-level researcher at Lancaster University. Her research focuses on the convergence of artificial intelligence with life sciences and biomedicine.