Extracting Tailings Ponds from High Spatial Resolution Remote Sensing Images using Improved YOLOv5 and SegFormer

Zhenhui Sun[1], Yunxiao Sun[1], Qingyan Meng[2,3,4*],
Tamás Jancsó[5*]

[1] School of Geology and Geomatics, Tianjin Chengjian University, Tianjin, China, 300384

[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, 100101

[3] University of Chinese Academy of Sciences, Beijing, China, 100049

[4] Key Laboratory of Earth Observation of Hainan Province, Hainan Aerospace Information Research Institute, Sanya, China, 572029

[5] Alba Regia Faculty, Obuda University, Budai út 45, 8000 Székesfehérvár, Hungary

e-mail: sunzh@tcu.edu.cn, sunyx@tcu.edu.cn, mengqy@radi.ac.cn, jancso.tamas@amk.uni-obuda.hu

Abstract: Dam failures in tailings ponds pose severe threats to nearby ecosystems, residents' lives, and property. Therefore, accurately and efficiently extracting information on tailings ponds is essential. Remote sensing technology has become a crucial tool for periodic and precise detection of these ponds. However, tailings ponds vary significantly in color, scale, and shape, often blending with their surroundings, which limits the effectiveness of traditional remote sensing methods. In this paper, we propose a framework for extracting tailings ponds from high-resolution remote sensing images using an improved YOLOv5 and SegFormer. Our improved YOLOv5 incorporates the coordinate attention (CA) and Transformer attention mechanisms into the C3 module of the backbone, creating new C3CA and C3TR modules that form a hybrid attention mechanism backbone. For the neck network, we build on YOLOv6's Bi-directional Concatenation (BiC) module, replacing the nearest-neighbor interpolation with transposed convolution, and designing a new BiCT module to create the BiC Transposed Convolution Path Aggregation Network (BiCTPAN). Following detection by the improved YOLOv5, SegFormer is used to accurately delineate tailings pond boundaries. The results show that the improved YOLOv5s achieves an mAP@0.5 of 90.10%, a 4.8% increase over the original YOLOv5s, with minimal impact on parameters and Floating-Point Operations per Second (FLOPs). The SegFormer model achieves an Intersection over Union (IoU) of 87.45% and an accuracy of 94.1%, demonstrating excellent extraction performance.

Keywords: tailings pond; YOLOv5s; attention mechanism; SegFormer

1 Introduction

A tailings pond serves as a storage facility for tailings and other industrial waste generated during the extraction and processing of mineral resources. It is primarily used for recovering mineral components remaining in tailings and recycling water. Tailings ponds can gradually become dangerous due to the high potential energy the harbor during the tailings storage process. The structural instability of tailings pond dams can lead to the rapid debris flow, resulting in extensive and large-scale damage. Moreover, tailings containing substantial quantities of heavy metals can cause severe pollution, potentially leading to disastrous consequences for the environment [1]. Due to mineral distribution and topographical considerations, tailings ponds are often located in remote mountainous areas or ecologically sensitive regions. Additionally, many enterprises neglect proper storage and disposal due to cost constraints, resulting in major accidents [2]. Hence, timely and efficient acquisition of information regarding the distribution of tailings ponds is crucial for effective supervision and emergency management of these facilities.

Tailings ponds are numerous and widely distributed. The traditional manual investigation method is time-consuming, laborious and limited by ground conditions, making it unable to meet the high timeliness requirements of tailings pond monitoring. Remote sensing technology offers several advantages, including extensive coverage and rapid data acquisition cycle. Consequently, it has become a pivotal technology for current tailings pond monitoring and identification. Traditional approaches can be categorized into three types: (1) visual interpretation method. Farrand et al. [3] used a constrained energy minimization technique to map the distribution of mine tailings in the Coeur d'Alene River Valley. Xiao et al. [4] used visual interpretation signs to identify and extract tailings pond information through human-computer interaction. (2) Index construction method. Ma et al. [5] employed Landsat 8 Operational Land Imager (OLI) data and a newly constructed ultra-low-grade iron index along with temperature information to accurately identify tailings. Hao et al. [6] introduced a tailings extraction model using various tailings indices based on iron-bearing minerals. (3) Classification identification method. Wu et al. [7] designed a support vector machine method for automatic detection of tailings ponds. Yu et al. [8] employed an object-oriented and random forest method for tailings pool identification.

The close integration of deep learning technology and remote sensing has spurred extensive exploration of tailings extraction methods based on deep learning technology. Lyu et al. [9] proposed a new deep learning-based framework for extracting tailings pond margins from high spatial resolution (HSR) remote sensing images by combining YOLOv4 and the random forest algorithm. Yan et al. [10] improved the Faster R-CNN deep learning object detection model by increasing the inputs from three true-color bands to four multispectral bands. Zhang et al. [11] proposed a Pseudo-Siamese Visual Geometry Group Encoder-Decoder network to achieve high accuracy in tailing pond extraction from VHR images. Wang et al. [12] proposed a fast tailings pond extraction method (Scene-Classification-Semantic-Segmentation, SC-SS) that couples scene classification and semantic segmentation models.

In summary, while there has been substantial progress in the research on tailings pond detection, traditional methods based on manually designed features are difficult to obtain accurate detection results. Although deep learning has enhanced the accuracy of tailings pond detection, the complex background, distinct variations in tailings pond characteristics, and an uneven and sparse spatial distribution pose significant challenges. Therefore, this paper proposes a framework for extracting tailings ponds from high-resolution remote sensing images using improved YOLOv5 and SegFormer. For sample set construction, model misdetection is reduced by adding negative samples. In improved YOLOv5, a novel hybrid attention backbone is constructed by integrating CA and Transformer attention mechanisms with the C3 module of YOLOv5 backbone. Furthermore, we incorporate the idea of the BiC module and use transposed convolution in the neck, resulting in the formation of the BiCTPAN neck. The SegFormer identifies the boundaries of the tailings ponds based on the detection results obtained from improved YOLOv5.

2 Study Area and Data

2.1 Study Area

Laiyuan County is located in the northwest of Baoding City, Hebei Province, as shown in Figure1. It has the highest density of tailings ponds in the Beijing-Tianjin-Hebei region, approximately 200/km^2, with various types and high potential risks [13]. Furthermore, the area is populated with ground objects that resemble tailings ponds, including reservoirs, exposed soil, barren rock, and man-made structures. These similar ground objects can significantly influence the precision of tailings pond extraction. Therefore, we selected Laiyuan County and its surrounding areas as the research area, which is of great significance for verifying the performance of the tailings ponds extraction method and meeting actual regulatory needs.

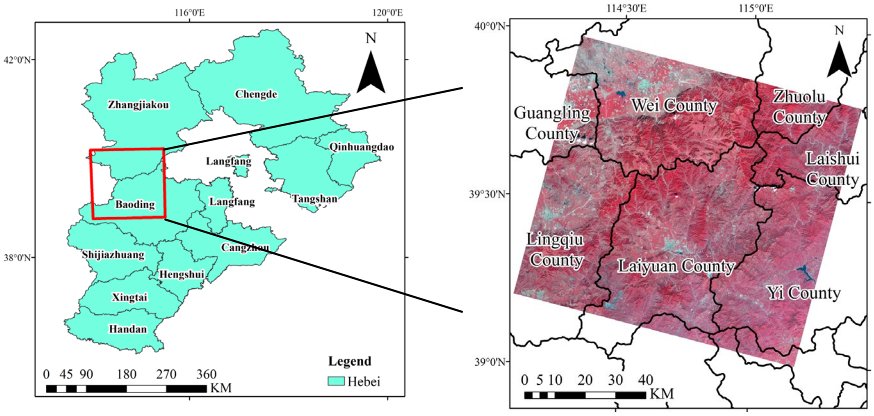


Figure 1
Study area(Image source: gaofen-6)

2.2 Data and Preprocessing

Gaofen (GF)-6 was successfully launched on June 2, 2018, and it boasts advanced imaging capabilities. It is equipped with a 2-meter panchromatic/8-meter multispectral high-resolution camera, offering a wide observation width of 90 kilometers. Additionally, it features a 16-meter multispectral medium-resolution wide-format camera with an impressive observation width of 800 kilometers. For this study, the images used were obtained by the multispectral high-resolution camera on September 6th, 2019, with cloud coverage measuring less than 10%.

The acquired data is at the L1A level and requires several preprocessing steps, including radiometric calibration, atmospheric correction, and orthorectification. Initially, the original data undergo radiometric calibration, using the GF-6 calibration parameters. Subsequently, the FLAASH model is applied for atmospheric correction. To rectify geometric distortions and enhance geometric accuracy, the atmospherically corrected data is orthorectified, relying on the rational polynomial coefficient file of the image and the corresponding digital elevation model data. Wang et al. [14] identified the GF-1 standard false color image as the optimal band combination for effectively identifying tailings ponds. Given the similarity in high spatial resolution camera parameters between GF-6 and GF-1, standard false-color images from GF-6 were employed for tailings pond extraction. It's worth noting that the original GF-6 image data is in 12-bit format, which has been converted to 8-bit for this study. The length and width of tailings ponds typically fall within the range of 50 to 3000 meters, and their external perimeters can span from 300 to 12600 meters, making them quite prominent in imagery [13]. Therefore, this study uses images with an 8-meter resolution for the detection of tailings ponds.

3 Methodology

The proposed framework for tailings pond extraction is shown in Figure 2. The framework can be summarized in the following steps: (1) Incorporate the Coordinate Attention (CA) and Transformer attention mechanisms into the YOLOv5 backbone to create the Hybrid Attention (HA) mechanism backbone. This enhancement bolsters the model's capability to extract features effectively. (2) Implement the concept of the BiC module from YOLOv6 into the neck of YOLOv5, resulting in the design of the BiPAN neck network. This modification improves the fusion of model features. (3) Use the improved YOLOv5 detection results as a foundation for employing SegFormer to extract the boundaries of the tailings ponds.

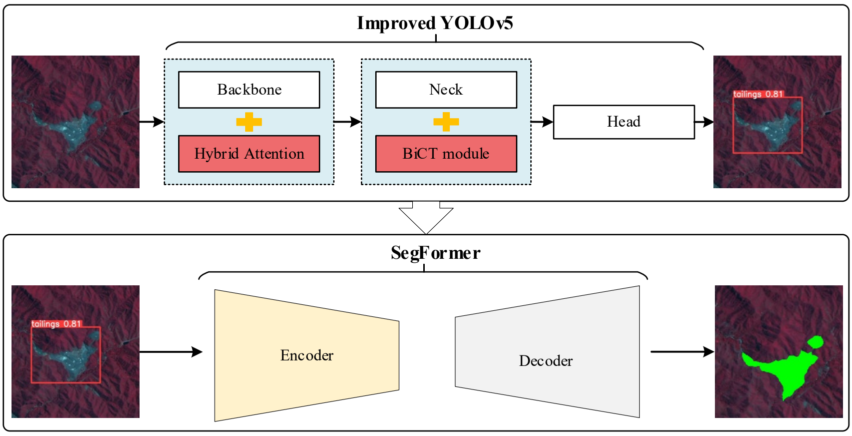


Figure 2
Flowchart for extraction of tailings ponds

3.1 Hybrid Attention Backbone

The introduction of attention mechanisms has been demonstrated to significantly enhance the performance of various computer vision tasks [15]. Among the most widely used attention mechanisms is the Squeeze and Excite (SE) attention [16]. However, SE only considers the attention in the channel dimension and ignores the information in the spatial dimension. The convolutional block attention module (CBAM) simultaneously pays attention to the channel and spatial dimension attention information [17]. CA outperforms other attention modules (e.g., SE, CBAM) by factorizing the 2D global pooling operations into two one-dimensional encoding processes [18]. The C3 module serves as a key module for YOLOv5 to learn more features. In the entire GF-6 image, a large number of small-sized tailings ponds are generally sparsely and non-uniformly distributed, making it difficult to distinguish them from the surrounding background, which makes tailings pond extraction challenging. The YOLOv5 with the C3 module cannot overcome this deficiency well because it lacks the ability to obtain global and contextual information [19], but the transformer can better integrate the semantic information of contextual features and global features and has a good recognition effect for sparse small targets with complex backgrounds [20] [21]. We enhance the C3 model by incorporating a transformer to form a C3TR module plugged into the backbone. Due to the high computational cost of the transformer, it is not feasible to incorporate the C3TR module into the network extensively. Therefore, the adoption of lightweight networks is necessary [22]. Considering the excellent performance of CA and its lightweight network structure, we concatenate the CA mechanism module with the C3 module to form the C3CA module. Multiple C3CA modules and a single C3TR module are integrated into the backbone to form a hybrid attention backbone. This enables the backbone to effectively capture long-range dependencies and global information in images without significantly increasing computational complexity, thus enhancing the extraction of meaningful features. The hybrid attention backbone is illustrated in Figure 3.

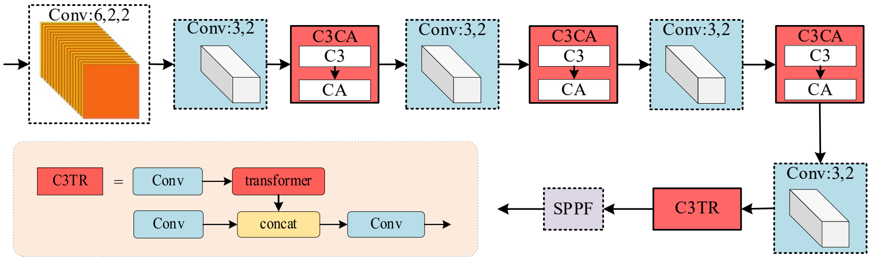


Figure 3
Hybrid attention backbone

3.2 BiCTPAN Neck

Multi-scale feature fusion is a crucial component of target detection. Feature Pyramid Network (FPN) merges low-level and high-level semantic features in a top-down manner to provide more accurate localization. YOLOv5 neck extends FPN by introducing a bottom-up path, facilitating the precise transmission of low-level feature information. However, YOLOv5 neck may not extract features comprehensively, leading to a decrease in target recognition accuracy [23]. To obtain more precise positioning information, this paper draws inspiration from YOLOv6 (v3.0) and incorporates the concept of its BiC module into the neck, resulting in the design of the BiCT module. This modification allows low-level features to participate more efficiently in multi-scale feature fusion, further enhancing the expressive capacity of the fused features. The BiCT module is depicted in Figure 4, where BiCT combines feature maps from three adjacent layers. In contrast, to YOLOv5's neck, there are two feature paths from the backbone, namely C_{i-1} and C_i, ensuring more comprehensive feature fusion from the backbone network. Additionally, the original upsampling method has been replaced with transposed convolution (ConvTranspose2d). BiCT is integrated into the YOLOv5 neck, forming BiCTPAN, as shown in Figure 4.

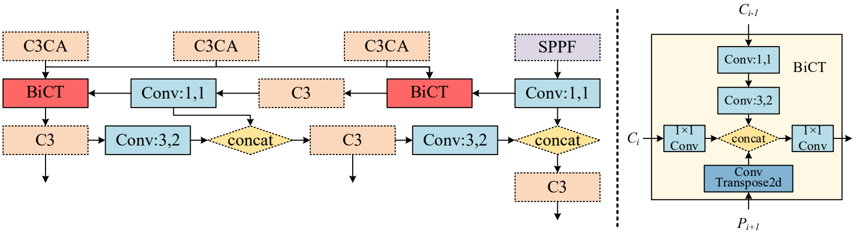


Figure 4
BiCTPAN neck

3.3 SegFormer

SegFormer is a Transformer-based semantic segmentation model [24], comprising an encoder and a decoder, as illustrated in Figure 5. The SegFormer encoder is equipped with four Transformer Blocks responsible for generating feature maps with different resolutions. These Transformer Blocks include Efficient Self-Attention (ESA), Mix-FFN, and Overlapped Patch Merging (OPM). The Vision Transformer (ViT) divides the input image into non-overlapping patches, disrupting the local continuity around these patches. The introduction of OPM effectively addresses this issue, enabling the extraction of large tailings ponds that have been split into different adjacent patches to still achieve excellent results. Self-attention plays a pivotal role in capturing the global aspect of the image but can be computationally intensive. ESA improves computational efficiency by reducing sequence length, making it a key component. Mix-FFN serves the purpose of adding location information to feature maps, contributing to the model's effectiveness. The decoder in SegFormer consists of Multilayer Perceptrons (MLP), which are lightweight yet efficient in capturing range information about the tailings reservoir. Firstly, multi-level features are fed into the MLP layer to normalize channel dimensions. Subsequently, the feature map is upsampled to 1/4 of the original image size and concatenated. Finally, an additional MLP layer is employed to aggregate the feature channels and perform classification prediction.

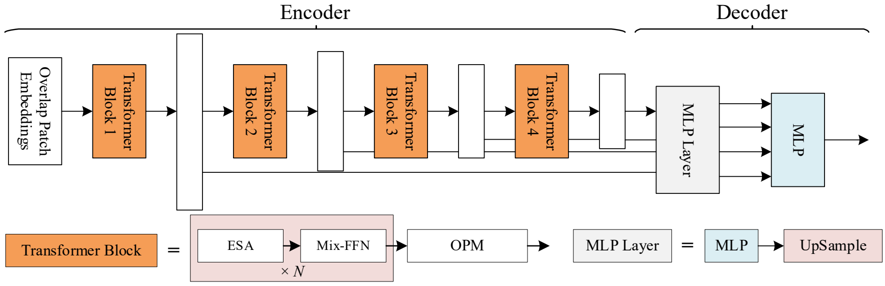


Figure 5
The architecture of SegFormer

3.4 Evaluation Methods

The Mean Average Precision (mAP) is a commonly used metric for assessing the performance and reliability of object detection models. In this paper, mAP@0.5 is employed to evaluate the performance of the YOLOv5 model in detecting tailings ponds. It represents the mAP value when the Intersection over Union (IoU) threshold is set to 0.5. IoU is the ratio of the intersection and union of two sets of real labels and predicted values for a specific category. SegFormer chooses IoU and accuracy (Acc) as evaluation metrics, where Acc represents the proportion of correctly predicted pixels out of the total pixels. Additionally, the article also includes parameters and Floating-Point Operations per Second (FLOPs) as indicators to measure the storage and computational resource requirements for model operation, respectively.

$$\text{Precision=True}_{\text{Positive}}/(\text{True}_{\text{Positive}}+\text{False}_{\text{Positive}}) \tag{1}$$

$$AP=\sum \text{Precision/(Total number of objects)} \tag{2}$$

$$\text{mAP=}\sum \text{Average precision/(Total number of classes)} \tag{3}$$

$$IoU = I(t)/U(t) = (\sum\nolimits_{n\in N} t_n * \Upsilon_n)/(t_n + \Upsilon_n - t_n * \Upsilon_n) \tag{4}$$

$$\text{Acc=(True}_{\text{Positive}}+\text{True}_{\text{negative}})/(\text{True}_{\text{Positive}}+\text{True}_{\text{negative}}+\text{False}_{\text{Positive}}+\text{False}_{\text{negative}}) \tag{5}$$

where $\text{True}_{\text{positive}}$ signifies that the predicted and actual class is positive, $\text{Talse}_{\text{positive}}$ means that the predicted class is positive and the actual class is negative, $\text{True}_{\text{negative}}$ means that the predicted class is negative and the actual class is positive, while in a $\text{False}_{\text{negative}}$, the predicted and actual class is negative. IoU defined by Equation (4) gives the ratio of intersection and union of the predicted bounding box and ground truth bounding box. t represents the probability outputs of pixel set n after filter by activation function in the GF-6 image; γ denotes the data set composed of ground truth bounding box.

3.5 Experimental Environment

The processor is Intel(R) Core(TM) i7-8750H at 3.80 GHz, memory 16G, GPU is NVIDIA RTX2070. The operating system is Windows 10, 64-bit, Cuda 11.3 and cuDNN 8.4.0. The depth learning framework is PyTorch 1.12.1.

The hyperparameters of YOLOv5 are set as follows: the training steps are 300 epochs; the warmup epoch and warmup momentum are respectively set as 3 and 0.8; the training and test batch size is 16. The optimization algorithm is a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01; the momentum and weight decay are respectively set as 0.937 and 0.0005.

The hyperparameter settings of the SegFormer model are as follows: due to limited GPU memory, the B1 model of SegFormer is used; the number of iterations is 8000; the loss function is cross-entropy loss; the batch size (batch size) is set to 4; the learning rate is initialized to 0.00006. The optimizer is AdamW; the weight decay coefficient (weight decay) is 0.01.

4 Results and Discussion

4.1 Sample Preparation

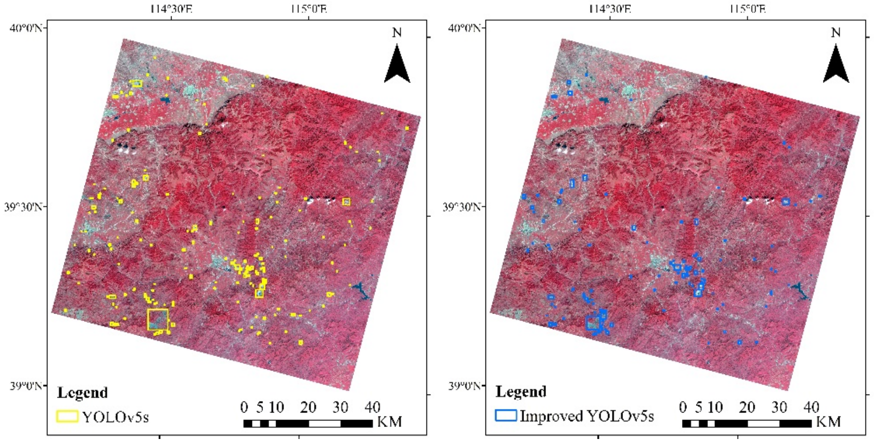
We labeled 950 tailings pond samples for training the improved YOLOv5 to detect tailings ponds. To mitigate the model's tendency for erroneous detection of similar ground objects and improve its generalization, we also labeled 275 negative samples. Of these, 90% was allocated for the training dataset, and the remaining 10% were reserved for the validation set. Considering the constraints of computing hardware, such as GPU limitations, the sample size was set to 500 × 500 pixels with a resolution of 8 meters. To ensure the fairness of the experiment, the same dataset was employed for all tailings pond detection experiments.

For SegFormer training, the sample image size was set to 600×600 to ensure coverage of tailings ponds. To obtain more targeted samples and achieve better extraction results, we initially extracted the center points from the detection frames produced by the improved YOLOv5 model. These center points were then used as the centers for generating the training samples. The dataset was divided into training and validation sets with a ratio of 0.85:0.15 for effective training and evaluation.

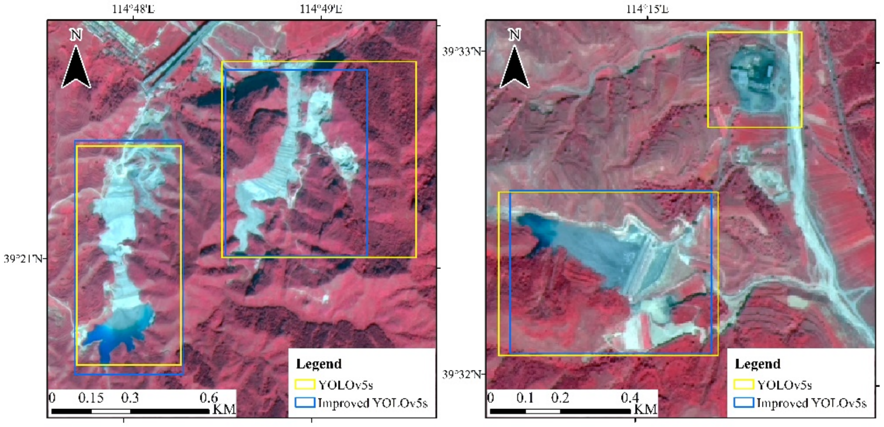
4.2 Results

4.2.1 Results of Improved YOLOv5

YOLOv5 is a widely used deep learning framework that comprises five network models of different sizes denoted as s, m, l, x, and n, representing various depths and widths of the network. In this study, we selected the YOLOv5s model size. Figure 6 illustrates the results of tailings pond detection using both YOLOv5s and the improved YOLOv5s. The yellow detection frames represent the results obtained by YOLOv5s, while the blue detection frames represent the results achieved with the improved YOLOv5s. Figure 6(a) and (b) are the results of using YOLOv5s and improved YOLOv5s to detect tailings ponds from the entire GF-6 image by combining overlapping slicing and a global non-maximum suppression algorithm [25]. Figures 6(c) and (d) depict results for specific local regions. Figures 6(c) and (d) showcase results for specific local regions. In Figure 6(c), it can be observed that both YOLOv5s and the improved YOLOv5s successfully detected the two tailings ponds in the image. But judging from the range of the detection frame, the improved YOLOv5s exhibits more accurate localization of the tailings ponds. Figure 6(d) shows an instance where YOLOv5s exhibits an error in detection. The ground object marked only with a yellow frame is a factory that is very similar to a tailings pond. In summary, the improved YOLOv5s proves to be more accurate in identifying tailings ponds and effectively reduces false detections.



(a) (b)



(c) (d)

Figure 6

The comparison results of YOLOv5s and improved YOLOv5s

Table 1 presents the quantitative results for YOLOv5s and the improved YOLOv5s. In comparison to YOLOv5s, the mAP@0.5 of the improved YOLOv5s increased by 4.80%, reaching 90.10%. Despite a slight increase in the number of parameters by 0.76 M and FLOPs by 0.4G, these increments are negligible when considering the advanced storage and computational resources. Therefore, the improved YOLOv5s achieves a significant boost in accuracy with virtually no increase in computational cost.

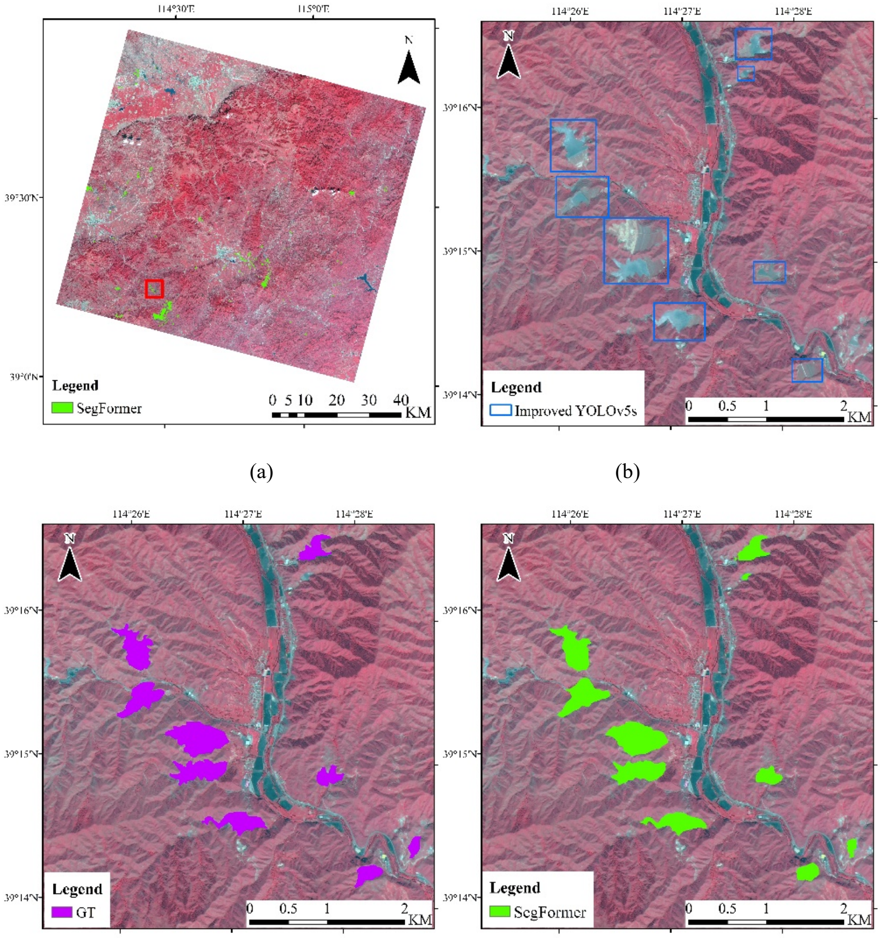
Table 1

Quantitative comparison results of YOLOv5s and improved YOLOv5s

models	parameters/M	mAP@0.5/%	FLOPs/G
YOLOv5s	7.02	85.30	15.80
Improved YOLOv5s	7.78	90.10	16.20

4.2.2 Results of SegFormer

Based on the detection results obtained using the improved YOLOv5s, SegFormer was employed to determine the extent of the tailings ponds. The extraction outcomes are presented in Figure 6. Figure 7(a) displays the extraction results of tailings ponds from the entire GF-6 image. To further illustrate the extraction performance of SegFormer, a specific local region was selected for display, as indicated by the red frame in Figure 7(a). Figure 7(b) shows the detection results based on the improved YOLOv5s, revealing that all tailings ponds in the region were successfully detected with accurate bounding frame positioning. In Figure 7(c), the manually labeled tailings ponds range information for this area is presented, while Figure 7(d) shows the outcomes of tailings ponds range extraction using the SegFormer. It can be seen that the tailings ponds detected by the improved YOLOv5s are all detected by the SegFormer. Compared with the manually labeled tailings ponds, the SegFormer model identifies Range information is very precise.



(a)

(b)

(c)

(d)

Figure 7

The comparison results of YOLOv5s and improved YOLOv5s

Table 2 presents the quantitative results for the SegFormer extraction of tailings ponds. It achieves an IoU of 87.45% and an Acc of 94.10%, demonstrating excellent extraction accuracy and reliability. Additionally, the model maintains a small number of parameters and FLOPs, resulting in minimal storage and computational resource consumption.

Table 2
The Quantitative comparison results of SegFormer

category	IoU/%	Acc/%	parameters /M	FLOPs/G
background	98.08	98.89	13.68	15.29
tailing ponds	87.62	94.28		

4.3 Discussion

4.3.1 Ablation Experiment

To assess the impact of different improvement strategies on the model, ablation experiments were conducted, and the results are presented in Table 3. According to Table 3, YOLOv5s with a hybrid attention backbone exhibits a 3.4% increase in mAP@0.5 compared to YOLOv5s, while maintaining a similar number of parameters and a reduction of 0.2G in FLOPs. In comparison, YOLOv5s with a BiCTPAN neck achieves a 3.1% increase in mAP@0.5 over YOLOv5s. However, this improvement comes with a slight increase of 0.16M parameters and an increase of 0.9G in FLOPs. The improved YOLOv5s demonstrates the highest improvement in mAP@0.5 compared to the baseline YOLOv5s. However, it also exhibits the most significant increase in parameters. Nevertheless, FLOPs are reduced by 0.5G when compared to YOLOv5s with BiCTPAN.

Table 3
Ablation study results

models	parameters /M	mAP@0.5/%	improvement over YOLOv5s /%	FLOPs/ G
YOLOv5s	7.02	85.30	--	15.80
+HA backbone	7.07	88.70	+3.40	16.00
+BiCTPAN neck	7.18	88.40	+3.10	16.70
improved YOLOv5s	7.78	90.10	+4.80	16.20

4.3.2 Comparison with Other Object Detection Methods

To demonstrate the effectiveness of the improved YOLOv5s in detecting tailings ponds on GF-6 images, this study compares the performance of our method with that of several other object detection methods, such as YOLOv8s, YOLOv5l, YOLT [26] and improvedv8s. The improved yolov8s has the same structure as the improved yolov5s. The improved yolov8s has the same structure as the improved yolov5s. Since YOLOv8 uses the C2f module instead of C3, the C3CA model is changed to the C2fCA module, which combines C2f and CA.

Table 4 compares the performance of various methods. Our approach and the improved YOLOv8s achieve the highest mAP@0.5, followed by the improved

YOLTv5s, YOLOv8s, and YOLOv5l. YOLOv8, as one of the latest models in the YOLO family, incorporates a new C2f module and a decoupled head, yielding excellent performance. YOLTv5, an evolution of YOLOv5, represents the fifth version of YOLT, and we selected the "s" (small) model size for comparison. Compared to YOLOv5s, YOLOv5l has a larger depth and layer channel multiplier, which generally provides better detection performance. Although the improved YOLOv8s also demonstrates strong results, it has a high parameter count and FLOPs. Our method is highly competitive in terms of parameter count and FLOPs, with only a slight increase of 0.72M parameters and 0.2G FLOPs compared to the best YOLT model.

Table 4
Experimental results of comparative experiments

models	parameters /M	mAP@0.5/%	FLOPs/G
YOLOv5l	46.11	87.60%	108.2
YOLOv8s	11.13	88.00%	28.60
YOLTv5s	7.06	88.60%	16.00
improved YOLOv8s	11.32	90.10%	29.40
Ours	7.78	90.10%	16.20

4.3.3 Attention Mechanism Comparison

To evaluate the effectiveness of the hybrid attention backbone, it was compared with backbones incorporating SE, CBAM, and CA attention mechanisms. SE and CBAM were integrated into the C3 module in a manner similar to C3CA resulting in the creation of C3SE and C3CBAM modules. The SE, CBAM, and CA attention mechanism backbones were constructed by replacing all C3 modules in the YOLOv5s backbone with C3SE, C3CBAM, and C3CA, respectively. Table 4 presents the comparison results for different attention backbones. The results demonstrate that the hybrid attention backbone achieves the highest mAP@0.5. Compared to the SE attention backbone network, mAP@0.5 increases by 2.2%. In comparison to the CBAM attention backbone network, mAP@0.5 increases by 2.4%. Finally, in comparison to the CA attention backbone network, mAP@0.5 increases by 1.4%. However, the changes in model parameters and FLOPs are relatively minor.

Table 4
Quantitative comparison results of YOLOv5s and improved YOLOv5s

models	parameters/M	mAP@0.5/%	FLOPs/G
YOLOv5s	7.02	85.30	15.80
SE attention backbone	7.06	86.50	15.80
CBAM attention backbone	7.06	86.30	15.80
CA attention backbone	7.05	87.30	15.80
Hybrid attention backbone	7.07	88.70	16.00

4.3.4 Model Generality

To verify the generalizability of the model, we selected a GF-6 image from Xuanhua District, Hebei Province. This image, captured by a high-resolution multispectral camera on October 21, 2019, has cloud coverage below 10%. After data preprocessing, we labeled 1,489 samples, with an equal split of positive and negative samples. The dataset was divided into training and validation sets at a 9:1 ratio, and each sample was 500×500 pixels. Training parameters remained unchanged. We also conducted ablation experiments on this new dataset, with results shown in Table 5. Table 5 demonstrates that each model component contributes to the accurate identification of tailings ponds. Compared to the original YOLOv5, our improved YOLOv5 achieved an mAP increase of +0.17, with minimal change in parameter count and FLOPs, indicating strong model generalizability.

Table 5

Ablation study results in different regions

models	parameters /M	mAP@0.5/%	improvement over YOLOv5s /%	FLOPs/ G
YOLOv5s	7.02	0.851	--	15.80
+HA backbone	7.07	0.855	+0.04	16.00
+BiCTPAN neck	7.18	0.859	+0.08	16.50
improved YOLOv5s	7.19	0.869	+0.18	16.60

4.3.5 Limitations and Future Work

While the improved YOLOv5s achieved enhanced detection results for tailings ponds, there are still instances of misidentification. Figure 8 illustrates some typical cases of model misidentification. These cases mainly involve artificial objects such as buildings, with a few instances of bare soil targets. The misidentifications occur due to their resemblance to tailings ponds in terms of shape and spectral characteristics. In addition, YOLOv5s takes 71.76 seconds to detect all tailings ponds from GF-6 images, while the improved YOLOv5s takes 108.98 seconds, which is 1.52 times longer than YOLOv5s. SegFormer also experienced extraction errors. Some error cases are shown in Figure 9. It can be seen that some artificial objects and bare soil are incorrectly extracted not only by the improved YOLOv5s, but also by SegFormer. We should not ignore that the samples for training SegFormer depend on the results of the improved YOLOv5s. Some samples missed by YOLOv5s are not considered, which may affect the extraction accuracy of SegFormer. Moreover, using two models also increases the running time of the extraction framework.

Additionally, several key challenges in identifying tailings ponds require attention. Tailings ponds in different regions exhibit significant variations in remote sensing images, including differences in brightness, hue, and scale. Due to diverse construction methods, these ponds may appear in various shapes, such as rectangles, triangles, circles, and irregular polygons. They are also situated against varied backgrounds, including vegetation, bare land, and sand. To improve the accuracy of tailings pond identification, it is essential to consider a wider range of features and integrate different types of data. Literature [28, 29] and our findings suggest combining of multiple attention mechanisms outperforms a single attention mechanism. Investigating how the integration of different attention mechanisms and their incorporation into various parts of YOLOv5s affects tailings pond detection is a promising research direction. Seasonal and sensor channel variations can limit the model's applicability, it may be beneficial to consider the introduction of a Generative Adversarial Network (GAN) to mitigate the differences in image hues acquired during different seasons and with different sensors.

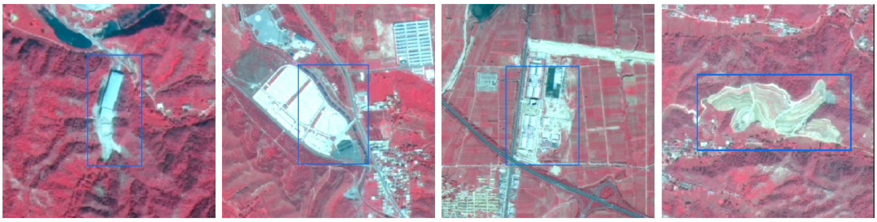


Figure 8

Some misidentified cases identified by improved YOLOv5s

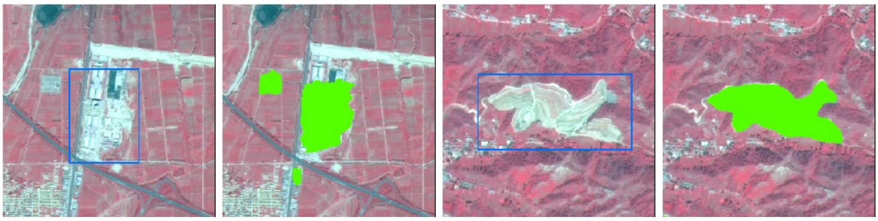


Figure 8

Some cases of incorrect extraction by SegFormer

Conclusions

In this paper, we propose a framework for extracting tailings ponds from GF-6 high spatial resolution remote sensing images using the improved YOLOv5 and SegFormer. The tailings pond dataset is generated based on the GF-6 high-resolution standard false color image and the strategy of incorporating negative samples. We introduced the C3CA and C3TR modules, which are formed by the CA and Transformer attention mechanisms, into the backbone network to construct a hybrid attention backbone. Additionally, we designed the BiCT module inspired by the BiC module in YOLOv6 to create the BiCTPAN neck.

SegFormer is employed to accurately delineate the range of tailings ponds based on improved YOLOv5 detection results. Our results demonstrate that compared to YOLOv5s, the improved YOLOv5s achieved a substantial 4.80% increase in mAP@0.5 with minimal additional computational cost. Furthermore, the hybrid attention backbone and BiCTPAN neck, which we designed, improved mAP@0.5 by 3.40% and 3.1%, respectively. The SegFormer model displayed remarkable accuracy in extracting tailings pond coverage, achieving an IoU of 87.45% and an Acc of 94.10%. Combining the improved YOLOv5 and SegFormer can yield high accuracy and stability in the extraction of tailings ponds. This method provides an effective tool for government agencies involved in tailings pond inventory and serves as a valuable reference for mine safety and environmental monitoring efforts.

Acknowledgement

This research was funded by the Bilateral Chinese-Hungarian Project(grant number 2019-2.1.11-TÉT-2020-00171); FY-3 Lot 03 Meteorological Satellite Engineering Ground Application System Ecological Monitoring and Assessment Application Project (Phase I)： ZQC-R22227; the Tianjin Municipal Education Commission Scientific Research Program (grant number 2021SK003); the Tianjin Educational Science Planning Project (grant number EHE210290); the National Natural Science Foundation of China (grant number 41971310); National Key Research and Development Program of China (grant number YFC3301602).

References

[1] Santamarina J. C., Torres-Cruz, L. A., Bachus, R. C.: Why coal ash and tailings dam disasters occur. Science, 2019; 364(6440) 526-528, DOI10.1126/science.aax1927

[2] Van Niekerk H. J., Viljoen M. J.: Causes and consequences of the Merriespruit and other tailings‐dam failures. Land degradation & development, 2005; 16(2) 201-212. DOI10.1002/ldr.681

[3] Farrand W. H., Harsanyi J. C.: Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho, through the use of a constrained energy minimization technique. Remote Sensing of Environment, 1997; 59(1): 64-76, https://doi.org/10.1016/S0034-4257(96)00080-6

[4] Xiao R., Shen W., Fu Z., Shi Y., Xiong W., Cao F.: The application of remote sensing in the environmental risk monitoring of tailings pond: A case study in Zhangjiakou area of China. In Earth Resources and Environmental Remote Sensing/GIS Applications III, 2012; pp. 332-339, DOI10.1117/12.964380

[5] Ma B., Chen Y., Zhang S., Li X.: Remote sensing extraction method of tailings ponds in ultra-low-grade iron mining area based on spectral characteristics and texture entropy. Entropy, 2018; 20(5): 345, https://doi.org/10.3390/e20050345

[6] Hao L., Zhang Z., Yang X.: Mine tailing extraction indexes and model using remote-sensing images in southeast Hubei Province. Environmental Earth Sciences, 2019; 219(78): 1-11, https://doi.org/10.1007/s12665-019-8439-1

[7] Wu X.: Image Extraction of Tailings Pond Guided by Artificial Intelligence Support Vector Machine. Wireless Communications & Mobile Computing (Online) 2022, https://doi.org/10.1155/2022/9406930

[8] Yu X., Zhang K., Zhang Y.: Land use classification of open-pit mine based on multi-scale segmentation and random forest model. Plos one, 2022; 17(2): e0263870, https://doi.org/10.1371/journal.pone.0263870.

[9] Lyu J., Hu Y., Ren S., Yao Y., Ding D., Guan Q., Tao L.: Extracting the tailings ponds from high spatial resolution remote sensing images by integrating a deep learning-based model. Remote Sensing, 2021: 13(4) 743, https://doi.org/10.3390/rs13040743

[10] Yan D., Zhang H., Li G., Li X., Lei H., Lu K., Zhu F.: Improved method to detect the tailings ponds from multispectral remote sensing images based on faster R-CNN and transfer learning. Remote Sensing, 2021; 14(1) 103, https://doi.org/10.3390/rs14010103

[11] Zhang C., Xing J., Li J., Du S., Qin Q.: A new method for the extraction of tailing ponds from very high-resolution remotely sensed images: PSVED. International Journal of Digital Earth, 2023; 16(1) 2681-2703, https://doi.org/10.1080/17538947.2023.2234338

[12] Wang P., Zhao H., Yang Z., Jin Q., Wu Y., Xia P., Meng L.: Fast Tailings Pond Mapping Exploiting Large Scene Remote Sensing Images by Coupling Scene Classification and Sematic Segmentation Models. Remote Sensing, 2023; 15(2) 327, https://doi.org/10.3390/rs15020327

[13] Li Q., Chen Z., Zhang B., Li B., Lu K., Lu L., Guo H.: Detection of tailings dams using high-resolution satellite imagery and a single shot multibox detector in the Jing–Jin–Ji region, China. Remote Sensing, 2020; 12(16) 2626, https://doi.org/10.3390/rs12162626

[14] Wang J., Cao L., Guo Y., Zhao L., Wu, B.: Feature analysis and information identification of the iron tailings by high−multispectral remote sensing. Journal of Yunnan University: Natural Sciences Edition, 2019; (5): 974-981, doi:10.7540/j.ynu.20180656

[15] Guo M., Xu T., Liu J., Liu Z., Jiang P., Mu T., Zhang S., Martin R., Cheng M., Hu, S.: Attention mechanisms in computer vision: A survey. Computational visual media, 2022; 8(3) 331-368, https://doi.org/10.1007/s41095-022-0271-y

[16] Hu J., Shen L., Sun G.: Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018; pp. 7132-7141, doi: 10.1109/TPAMI.2019.2913372

[17] Woo S., Park J., Lee J. Y., Kweon, I. S.: Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), 2018; pp. 3-19, https://doi.org/10.1007/978-3-030-01234-2_1

[18] Hou Q., Zhou D., Feng J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021; pp. 13713-13722, doi: 10.1109/CVPR46437.2021.01350

[19] Liu Y., He G., Wang Z., Li W., Huang H.: NRT-YOLO: Improved YOLOv5 Based on Nested Residual Transformer for Tiny Remote Sensing Object Detection. Sensors, 2022; 22(13): 4953, https://doi.org/10.3390/s22134953

[20] Zhu X., Lyu S., Wang X., Zhao Q.: TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF international conference on computer vision, 2021; pp. 2778-2788, doi: 10.1109/ICCVW54120.2021.00312

[21] Yu Y., Zhao J., Gong Q., Huang C., Zheng G., Ma J.: Real-time underwater maritime object detection in side-scan sonar images based on transformer-YOLOv5. Remote Sensing, 2021; 13(18) 3555, https://doi.org/10.3390/rs13183555

[22] Zhao S., Kang F., Li J.: Concrete dam damage detection and localisation based on YOLOv5s-HSC and photogrammetric 3D reconstruction. Automation in Construction, 2022; 143, 104555, https://doi.org/10.1016/j.autcon.2022.104555

[23] Yang J., Li H., Du Y., Mao Y., Liu Q.: Lightweight Object Detection Algorithm Based on Improved YOLOv5s, Electronics Optics & Control, 2023; 30(02):24-30, doi:10.3969/j.issn.1671-637X.2023.02.005

[24] Xie E., Wang W., Yu Z., Anandkumar A., Alvarez J. M., Luo P.; SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 2021; 34, 12077-12090, https://doi.org/10.48550/arXiv.2105.15203

[25] Sun Z., Li P., Meng Q., Sun Y., Bi Y.: An Improved YOLOv5 Method to Detect Tailings Ponds from High-Resolution Remote Sensing Images. Remote Sensing, 2023; 15(7) 1796, https://doi.org/10.3390/rs15071796

[26] Van Etten, A. You only look twice: Rapid multi-scale object detection in satellite imagery. arXiv 2018, arXiv:1805.09512

[27] Károly, A. I., Fullér, R., Galambos, P.. Unsupervised clustering for deep learning: A tutorial survey. Acta Polytechnica Hungarica, 2018;15(8) 29-53

[28] Deng T., Liu X., Mao G.: Improved YOLOv5 based on hybrid domain attention for small object detection in optical remote sensing images. Electronics, 2022; 11(17) 2657, https://doi.org/10.3390/electronics11172657

[29] Zhu L., Geng X., Li Z., Liu C.: Improving YOLOv5 with attention mechanism for detecting boulders from planetary images. Remote Sensing, 2021; 13(18), 3776, https://doi.org/10.3390/rs13183776