Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065339

ARTICLE

An Ochotona Curzoniae Object Detection Model Based on Feature Fusion with SCConv Attention Mechanism

Haiyan Chen* **and Rong Li**

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China

*Corresponding Author: Haiyan Chen. Email: chenhaiyan@sina.com

Received: 10 March 2025; Accepted: 19 June 2025; Published: 30 July 2025

ABSTRACT: The detection of Ochotona Curzoniae serves as a fundamental component for estimating the population size of this species and for analyzing the dynamics of its population fluctuations. In natural environments, the pixels representing Ochotona Curzoniae constitute a small fraction of the total pixels, and their distinguishing features are often subtle, complicating the target detection process. To effectively extract the characteristics of these small targets, a feature fusion approach that utilizes up-sampling and channel integration from various layers within a CNN can significantly enhance the representation of target features, ultimately improving detection accuracy. However, the top-down fusion of features from different layers may lead to information duplication and semantic bias, resulting in redundancy and high-frequency noise. To address the challenges of information redundancy and high-frequency noise during the feature fusion process in CNN, we have developed a target detection model for Ochotona Curzoniae. This model is based on a spatial-channel reconfiguration convolutional (SCConv) attentional mechanism and feature fusion (FFBCA), integrated with the Faster R-CNN framework. It consists of a feature extraction network, an attention mechanism-based feature fusion module, and a jump residual connection fusion module. Initially, we designed a dual attention mechanism feature fusion module that employs spatial-channel reconstruction convolution. In the spatial dimension, the attention mechanism adopts a separation-reconstruction approach, calculating a weight matrix for the spatial information within the feature map through group normalization. This process directs the model to concentrate on feature information assigned varying weights, thereby reducing redundancy during feature fusion. In the channel dimension, the attention mechanism utilizes a partition-transpose-fusion method, segmenting the input feature map into high-noise and low-noise components based on the variance of the feature information. The high-noise segment is processed through a low-pass filter constructed from pointwise convolution (PWC) to eliminate some high-frequency noise, while the low-noise segment employs a bottleneck structure with global average pooling (GAP) to generate a weight matrix that emphasizes the significance of channel dimension feature information. This approach diminishes the model's focus on low-weight feature information, thereby preserving low-frequency semantic information while reducing information redundancy. Furthermore, we have developed a novel feature extraction network, ResNeXt-S, by integrating the Sim attention mechanism into ResNeXt50. This configuration assigns three-dimensional attention weights to each position within the feature map, thereby enhancing the local feature information of small targets while reducing background noise. Finally, we constructed a jump residual connection fusion module to minimize the loss of high-level semantic information during the feature fusion process. Experiments on Ochotona Curzoniae target detection on the Ochotona Curzoniae dataset show that the detection accuracy of the model in this paper is 92.3%, which is higher than that of FSSD512 (84.6%), TDFSSD512 (81.3%), FPN (86.5%), FFBAM (88.5%), Faster R-CNN (89.6%), and SSD512 (88.6%) detection accuracies.

KEYWORDS: Ochotona curzoniae; target detection; SCConv attention; feature fusion

1 Introduction

Alpine meadow ecosystems are unique ecosystems distributed in cold regions at high altitudes (such as the Tibetan Plateau), playing a crucial role in biodiversity, climate regulation, water conservation and soil preservation [1]. Ochotona Curzoniae is a small burrowing mammal found on the Tibetan Plateau, with a significant impact on vegetation, soil properties, biodiversity, and grassland degradation in alpine meadow ecosystems [2]. The burrows dug by Ochotona Curzoniae improve the soil structure of alpine meadow ecosystems, increasing soil permeability and the organic matter and moisture content of the soil. However, when the population density of Ochotona Curzoniae exceeds a certain threshold, a large number of pikes excessively graze on pasture grasses, dig holes, destroy the soil structure of the pasture, and accelerate grassland degradation, leading to rodent infestation [3]. Therefore, to prevent Ochotona Curzoniae from damaging the alpine meadow ecosystem, it is essential to monitor their population density and take appropriate ecological control measures to maintain the population within a reasonable range.

Conventional techniques for assessing the population density of Ochotona Curzoniae, such as the burrow method [4] and the trapping method [5] within quadrants comprehensive population data. However, these methods are characterized by significant time and labor demands, making them unsuitable for continuous long-term monitoring of this species. Rapid advancements in video surveillance and hardware technologies [6] enable the integration of intelligent monitoring systems and computer vision into the assessment of Ochotona Curzoniae population density. This innovative approach not only facilitates sustained long-term monitoring but also broadens the scope of monitoring efforts [7]. The identification of Ochotona Curzoniae individuals is a critical component of computer vision methodologies aimed at estimating population sizes and analyzing fluctuations in population density. Consequently, the successful implementation of computer vision-based monitoring for Ochotona Curzoniae population density requires precise target detection capabilities.

The small size, low pixel proportion, and insignificant features of Ochotona Curzoniae in natural scene images increase the difficulty of target detection. Compared to traditional methods, target detection based on Deep Convolutional Neural Networks (DCNN) not only achieves higher detection accuracy but also significantly saves financial, material, and human resources [8]. Therefore, it has been applied to the detection of Ochotona Curzoniae targets. Taking advantage of the significant advantages of DCNN in feature extraction, many DCNN-based target detection models that balance detection performance and efficiency have emerged in recent years [9,10]. Faster R-CNN is one of the most representative detection models [11]. In the target detection process, it first uses a feature extraction network to obtain feature maps and then applies Non-Maximum Suppression (NMS) to filter candidate boxes generated by the Region Proposal Network (RPN), keeping the regions of interest. These regions are then passed into Fast R-CNN for location and class judgement to perform target detection. However, when using Faster R-CNN for target detection, repeated pooling operations can lead to loss of target feature information, causing missed detections and false positives [12]. This is especially problematic for small targets such as Ochotona Curzoniae, where repeated pooling exacerbates the loss of feature information, resulting in a significant number of missed and false detections.

Researchers have conducted comprehensive investigations aimed at enhancing the feature extraction capabilities of DCNN-based target detection algorithms. Their findings indicate that high-level features extracted by DCNNs possess greater semantic information, which is advantageous for target classification; however, these features exhibit lower resolution and diminished detail-capturing abilities [13]. Conversely, low-level features are characterized by higher resolution and contain more positional and detailed information that is beneficial for target localization. However, they are associated with lower semantic information and increased noise due to the reduced number of convolutional layers [14]. Consequently, the integration of information across various feature layers to enrich the representation of target features has emerged

as a pivotal area of research.To address this challenge, Li and Zhou [15] introduced a lightweight feature fusion module within the Single Shot MultiBox Detector (SSD), proposing an end-to-end feature fusion and feature-enhanced SSD (FSSD) target detection algorithm. This feature fusion module concatenates features from various layers and scales, subsequently generating a new feature pyramid through subsampling. The FSSD achieved a mAP of 82.7% on the Pascal VOC 2007 dataset, with a detection speed of 65.8 frames per second (FPS). Similarly, Pan et al. [16] developed a top-down feature fusion module within the SSD target detection network to improve performance in detecting targets of varying scales. The core of the Top-Down Feature Fusion SSD (TDFSSD) lies in its iterative fusion of high-level features, which encapsulate semantic information, and low-level features, which convey boundary information. TDFSSD attained an mAP of 81.7% on the Pascal VOC 2007 dataset, 80.1% on the Pascal VOC 2012 dataset, and an mAP of 17.2% for small targets on the COCO test set.

Furthermore, Lin et al. [17] proposed a Feature Pyramid Network (FPN) for multiscale target detection. The FPN integrates high-level semantic information with low-level detail through both bottom-up and top-down pathways, while also incorporating lateral connections to merge upsampled deep feature maps with shallow feature maps of corresponding scales via convolution. This approach enhances the semantic richness of the feature maps. Additionally, Chen et al. [18] utilized feature maps extracted by ResNet101 within the Faster R-CNN target detection model, generating a feature pyramid network through the Feature Fusion Block Attention Module (FFBAM) and producing multiscale feature information via a top-down fusion approach. This method achieved a mean Average Precision (mAP) of 82.5% for small target detection on the NWPU VHR-10 dataset, thereby improving detection performance for small targets.

The above literature indicates that integrating high-level feature information with low-level feature information and introducing low-level semantic information into high-level features can enhance detail perception while strengthening semantic understanding, thereby improving the detection accuracy of DCNN for small targets. However, the feature fusion process typically employs methods based on channel concatenation or element-wise multiplication. These methods not only lead to redundant feature extraction and information redundancy, but also amplify high-frequency noise in feature maps, resulting in aliasing effects [19,20]. Such aliasing effects not only reduce the detection accuracy for small targets but also, due to the differences in semantic information between different layers of feature maps, top-down layer-by-layer feature fusion can cause the loss of semantic information from upper-level features, leading to insufficient feature fusion [21].

To mitigate the challenges associated with information redundancy and high-frequency noise that arise during feature fusion, researchers have identified that implementing attention mechanisms can effectively address these issues. For instance, Zhao et al. [22] introduced a dynamic Spatial Transformers (ST) attention mechanism, which overcomes the limitations of traditional convolutional neural networks(CNN) in extracting features from images of varying scales. This mechanism achieves this by calculating the parameters necessary for the spatial transformation of each pixel within the feature map corresponding to a specific region of the input image. However, the ST mechanism primarily focuses on the spatial dimension of the feature map, neglecting the semantic information disparities among channels, which consequently limits its ability to mitigate aliasing effects. In related study, Hu et al. [23] implemented the Squeeze-and-Excitation (SE) mechanism after feature fusion to learn the interdependencies among different feature map channels within the network. This approach recalibrates the features to determine the significance of each feature channel, thereby enhancing relevant features while reducing redundant features and noise. However, the SE mechanism primarily operates in the channel dimension, which limits its effectiveness in suppressing redundant information within the network.

To address the limitations of previous attention mechanisms that focus exclusively on either spatial or channel dimensions, Woo et al. [24] proposed the Convolutional Block Attention Module (CBAM). This module integrates both channel and spatial information by capturing relationships between channels through global average pooling and global max pooling, while simultaneously evaluating the significance of various positions on the feature map. CBAM enhances salient features in the spatial dimension, thereby achieving attention to critical information across both channel and spatial dimensions, and effectively suppressing irrelevant information.

The existing literature indicates that while attention mechanisms can significantly reduce information redundancy and high-frequency noise during feature fusion, their effectiveness is limited when they focus solely on either spatial or channel dimensions. The integration of spatial-channel attention mechanisms has the potential to further decrease information redundancy and noise interference, thereby improving the model's detection accuracy [25].

To reduce redundant information and high-frequency noise during the feature fusion process, thereby improving the accuracy of the Faster R-CNN model for small target detection, this study proposes FFBCA. This approach aims to achieve precise detection of small targets. The primary innovations and contributions of this research are outlined as follows:

- To enhance the capacity of the shallower backbone network for localized feature extraction in images, this study proposes a novel network architecture called ResNeXt-S. This architecture integrates the Sim attention mechanism within the ResNeXt50 feature extraction framework, thereby improving the network's ability to capture local features within the feature map.
- To mitigate information redundancy and high-frequency noise arising from the fusion of features extracted by ResNeXt-S, a dual-attention mechanism based on spatial-channel reconfiguration convolution has been developed. The attention mechanism in the spatial dimension employs a segregation-reconstruction approach, which guides the model to reduce its focus on redundant information by generating a weight map characterized by local smoothing capabilities and parameter factors. Conversely, the attention mechanism in the channel dimension utilizes a segmentation-transposition-convolution strategy. In the spatial dimension, the attention mechanism incorporates a segmentation-transposition-fusion technique, while the bottleneck structure, established through low-pass filtering and convolution of Global Average Pooling (GAP), effectively reduces the information redundancy generated during the feature fusion process and simultaneously diminishes high-frequency noise.
- We designed a Skip Residual Connection Fusion (SRCF) approach for minimizing the loss of high-level information during feature fusion.

This paper is structured as follows: the initial section examines the challenges associated with the detection of Ochotona Curzoniae and reviews existing target detection methodologies that utilize feature fusion and attention mechanisms. The subsequent section delineates the architecture of the proposed spatial-channel reconfiguration convolutional attention mechanism for feature fusion in the Ochotona Curzoniae target detection model, along with a discussion of the key enhancement modules. The third section outlines the experimental framework, which includes the dataset and parameters employed, as well as the performance results of the proposed model on both the Ochotona Curzoniae dataset and the NWPU VHR-10 dataset. The fourth section presents comprehensive experiments aimed at assessing the efficacy of the improved modules. Finally, Section V concludes with a summary of the contributions of this paper and its overall organization.

2 Target Detection Model

The overall structure of the target detection model constructed in this paper is shown in Fig. 1. The backbone network is responsible for extracting the features of the regions of interest in the input image. The RPN is responsible for extracting candidate boxes from the shared feature map. Fast RCNN is responsible for regressing the size and position of the candidate boxes and determining the target category. Due to the small proportion of information for small targets in the image, a shallow network, ResNeXt50, is used for feature extraction. To enhance the extracted feature details, Sim attention is introduced into ResNeXt50 to build a new network: ResNeXt-S. To extract finer-grained feature information of small targets, reduce information redundancy and high-frequency noise during feature fusion, and prevent feature loss, two structures, FFBCA and SRCF, are designed. In Fig. 1, the different layer features extracted by the ResNeXt-S network are the features used for detection after fusion. FFBCA is the feature fusion based on SCConv attention mechanism designed in this paper. The red lines in Fig. 1 represent the SRCF method designed in this paper. Finally, the features extracted by the backbone network of this model are sent to the RPN and Fast RCNN for classification and regression to determine the final category and position information. It should be noted that this paper only improves the backbone network of Faster R-CNN, and the other parts are consistent with the original structure [26]. Therefore, only the constructed ResNeXt-S, FFBCA, and SRCF are explained here.

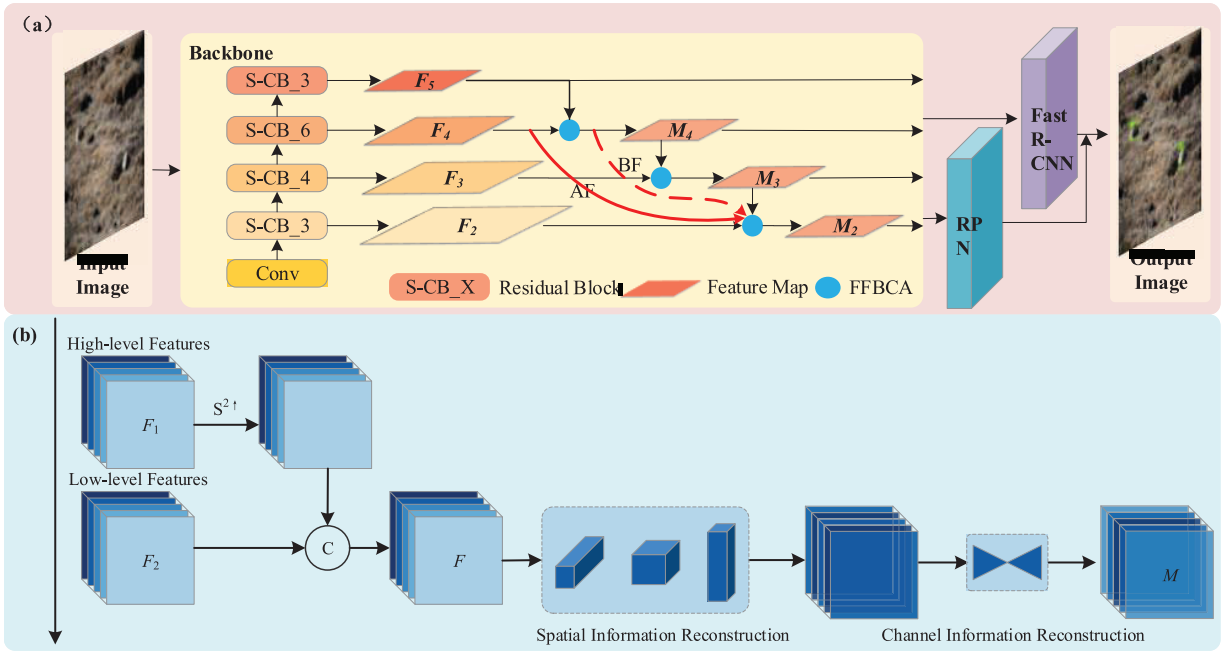

Figure 1: (**a**): The overall structure of the target detection model constructed in this paper. (**b**): Feature fusion based on SCConv attention mechanism

2.1 ResNeXt-S

In target detection, deepening the backbone network helps the network learn more feature information. However, because small targets occupy less space in the image, using deeper networks leads to a gradual reduction in the resolution of the feature maps, which decreases the ability to capture feature information for small targets [27]. Therefore, in this study, ResNeXt50 with 16 layers is selected as the feature extraction network. To enhance the network's ability to capture local features in the feature map, a Simple Attention Module (SimAM) is introduced in the original ResNeXt, creating a new feature extraction network:

ResNeXt-S. To study the impact of the position where the attention mechanism is introduced on detection accuracy, three different forms of introduction are constructed as shown in Fig. 2. Through experimental analysis, the feature extraction network is constructed using Fig. 2a.

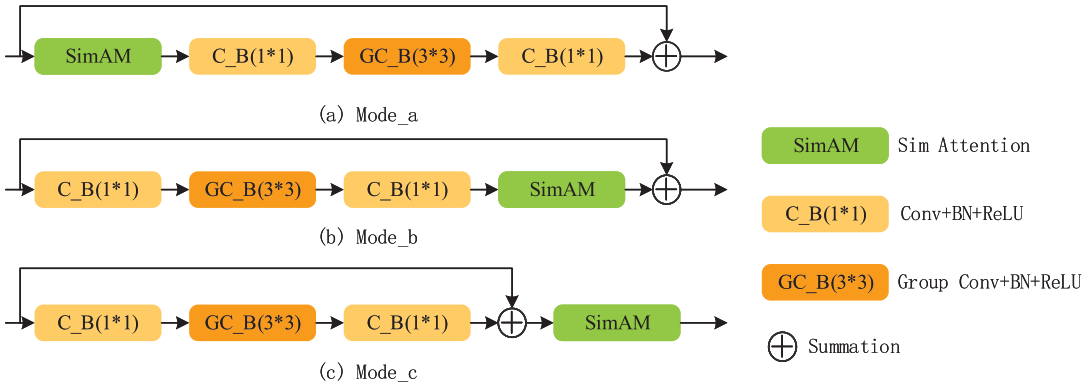

Figure 2: Sim-CB approach

Define the input feature $X \in R^{C\times W\times H}$, then the entire process of Fig. 2a can be represented by Eq. (1), where SimAM is the attention mechanism. $C\_B_{1\times1}$ represents applying a 1 × 1 convolution, batch normalization, and ReLU activation function to the input feature; $GC\_B_{3\times3}$ represents applying a 3 × 3 group convolution, batch normalization, and ReLU activation function to the input feature; ⊕ represents concatenation along the channel dimension, and the formulas for calculating $C\_B_{1\times1}$ and $GC\_B_{3\times3}$ are given in Eqs. (2) and (3).

$$\tilde{X} = C\_B_{1\times1}(GC\_B_{3\times3}\{C\_B_{1\times1}[SimAM(X)]\}) \oplus X \tag{1}$$

$$C\_B_{1\times1}(X) = \mathcal{R}\{\mathcal{BN}[Conv_{1\times1}(X)]\} \tag{2}$$

$$GC\_B_{3\times3}(X) = \mathcal{R}\{\mathcal{BN}[GConv_{3\times3}(X)]\} \tag{3}$$

The structure of SimAM is shown in Fig. 3. In the figure, GAP represents Global Average Pooling; Q is the attention weight matrix; σ represents the Sigmoid activation function; and B represents the Hadamard product. SimAttention evaluates the importance of each neuron in the network by introducing an energy function, deriving the three-dimensional attention weights for each neuron, thus capturing important information from the feature map more comprehensively.

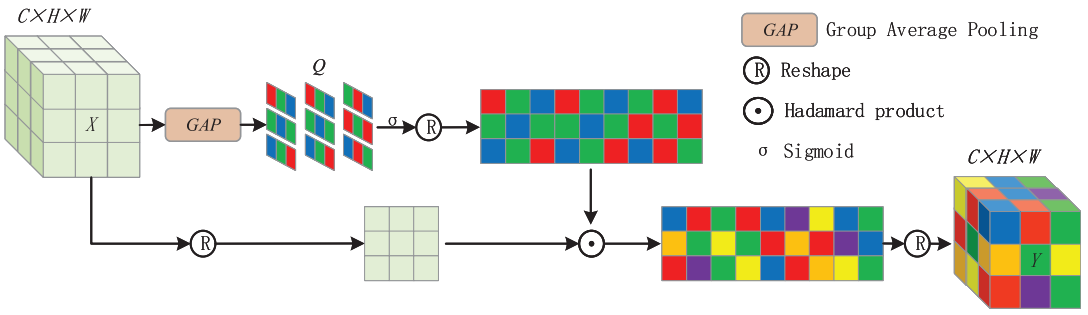

Figure 3: SimAM

First, GAP is used to calculate the attention weight for each pixel in the feature map, obtaining a three-dimensional weight matrix Q for the entire feature map. The specific calculation formula is shown in Eq. (4), where ξ is the normalization constant, $S(f_k, f_l)$ represents the similarity between the k-th and l-th pixels in the feature map, $l \in N_k$ (N_k represents the set of neighboring pixels of the k-th pixel), and q_k represents the attention weight generated for the k-th pixel in the feature map. Then, the input feature $X \in R^{C\times W\times H}$ is reshaped into $X \in R^{C\times QW\times QH}$; the three-dimensional weight matrix Q is first activated with Sigmoid and then reshaped into $Q \in R^{C\times W\times H}$. The reshaped X and Q are combined using the Hadamard product and reshaped to obtain a new feature, denoted as $Y \in R^{C\times W\times H}$. The entire computation process is shown in Eq. (5).

$$Q \in q_k = \frac{1}{\xi}\sum S(f_k, f_l),\ (k, l = 1, 2, \cdots, C) \tag{4}$$

$$Y = \text{Reshape}[\sigma(Q) \odot X] \tag{5}$$

In this paper, ResNeXt-S combines ResNeXt50 and Sim attention mechanism to enhance the feature extraction capability of small targets through group convolution and residual concatenation. ResNeXt50 utilizes group convolution to capture the different attributes of small targets, avoiding the limitation of a single convolution kernel in the expression of feature diversity. Sim attention mechanism enhances the discriminative features by adjusting the 3D weights to suppress redundant responses and improves the model's ability to capture small target feature information.

2.2 Feature Fusion Based on SCConv Attention Mechanism

In order to reduce the information redundancy and high-frequency noise generated in the process of feature fusion, this paper designs a feature fusion module based on the spatial-channel reconstruction convolutional attention mechanism, which firstly fuses the high-level features extracted by ResNeXt-S in Section 2.1 with the low-level features. Then based on the spatial-channel reconstruction convolutional attention mechanism in the spatial dimension through the group normalization to calculate the weight matrix of the spatial information in the feature map, through the weight threshold to distinguish between the high information content and low information content regions, to retain the regions with high information content, to reduce the sensitivity of the model to the low information content regions, so as to reduce the model's dependence on the redundant information; in the channel dimension in accordance with the variance of the feature information content of the feature map is partitioned into high noise and low noise parts, using the weights learned from PWC to judge the importance of features for the high noise part, thus suppressing high frequency noise, and using GAP to generate the channel weight map for the low noise part, which suppresses redundancy while retaining low frequency semantic information.

The structure of FFBCA is shown in Fig. 1b, where F_1 and F_2 represent the input high-level and low-level features, respectively, and $S^{2\uparrow}$ represents 2x upsampling. First, the input high-level feature F_1 is upsampled by a factor of 2 to the spatial size of the low-level feature F_2. Then, the upsampling feature is concatenated with the low-level feature along the channel dimension to obtain feature map F. Thereafter, spatial information reconstruction (SIR) and channel information reconstruction (CIR) are performed on feature map F to obtain the new feature map M, which not only retains the semantic information of high-level features but also integrates the detailed information of low-level features, providing a richer and more robust feature representation for subsequent object detection tasks.

SIR adopts a separation-reconstruction method to focus on the semantic redundant information between adjacent pixels in different feature spaces, reducing noise effects and enhancing the effective

information in the feature map. The structure of SIR is shown in Fig. 4. Let the input feature be defined as $F \in R^{B\times C\times W\times H}$. First, apply Group Normalization (GN) to estimate the spatial information in different feature maps, as shown in Eq. (6). In Eq. (6), μ and σ^2 represents the mean and variance, with its calculation formula given in Eq. (7). α and β is the learnable scaling factor, mainly used to characterize the sufficiency of spatial information. A larger α indicates a greater difference between adjacent pixels, i.e., richer spatial information. ε is a very small constant (set to 0.001) mainly used to ensure network stability.

$$GN(F) = \alpha \frac{F-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{6}$$

$$\begin{cases} \mu = \dfrac{\sum_{i=1}^{C} f_i}{C} \\ \sigma^2 = \dfrac{\sum_{i=1}^{C}(f_i-\mu)^2}{C} \end{cases} \tag{7}$$

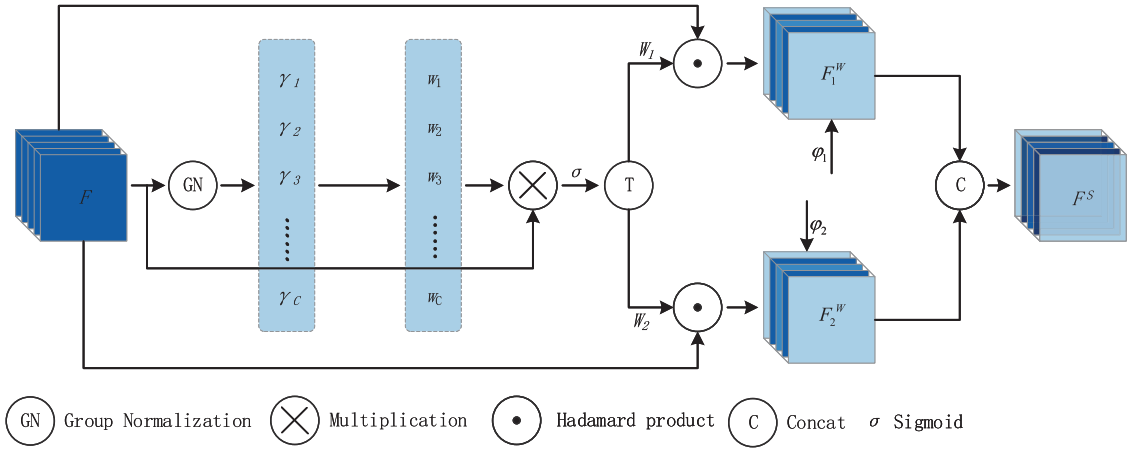

Figure 4: Spatial information reconstruction

After applying GN, F is aggregated into $GN(F)$ to obtain the spatial information of the feature. A weight factor W is then used to measure the importance of the spatial information. Suppose the spatial information of the i-th pixel in the feature map after GN operation is γ_i, the corresponding weight factor for this point can be calculated using Eq. (8), where γ_i represents the spatial information of the j-th pixel adjacent to i-th. The final weight factor matrix W, which measures the importance of each spatial dimension information in the input feature map F, is obtained as shown in Eq. (9). In Eq. (9), W_i and $GN(X_i)$ represent the weight factor corresponding to the i-th pixel in the feature map, the spatial information obtained after GN operation; σ represents the Sigmoid activation function; and $\otimes$ represents matrix multiplication. After obtaining W, it is first mapped to (0, 1), then a threshold T is applied to form a gate, dividing W into W_1 and W_2 according to the importance of the information. These are then used for Hadamard product with the input feature map F: ${F_1}^W = F \odot W_1$, ${F_2}^W = F \odot W_2$, ${F_1}^W$ contains more spatial information, while ${F_2}^W$ contains less spatial information (with a large amount of spatial information redundancy). In [28], cross-reconstruction is used to add features with more information to those with less, generating features with more information. In this paper, two constants are introduced to reduce the number of parameters and measure the importance of

feature information, denoted as φ: $F^S = \sigma\{Cat[\varphi_1 \odot {F_1}^W, \varphi_2 \odot {F_2}^W]\}$, φ_1 and φ_2, which are constants used to evaluate the degree of importance: $\varphi_1 + \varphi_2 = 1$.

$$w_i = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j}, (1 \le i \le H, 1 \le j \le W) \tag{8}$$

$$W \in \{W_i\} = \sigma(F \otimes GN(X_i)), (1 \le i \le H) \tag{9}$$

CIR adopts a segmentation-transposition-fusion method to reduce the redundant information between different channels, thereby suppressing information redundancy and noise effects, further enhancing the feature representation ability. The structure of CIR is shown in Fig. 5. In Fig. 5, GWC represents grouped convolution, PWC represents pointwise convolution, GAP represents global average pooling, and σ represents the Sigmoid activation function. When performing channel information reconstruction, the input feature $F \in R^{B\times C\times W\times H}$ is first split into two parts according to the number of channels λ using a hyperparameter: ${F_1}^S \in R^{B\times \lambda C\times W\times H}$, ${F_2}^S \in R^{B\times (1-\lambda)C\times W\times H}$, $\lambda \in (0,1)$. Since ${F_1}^S$ contains rich channel information, GWC is first applied to extract feature information. However, GWC cuts off the information flow between channels, so PWC is introduced to compensate for the loss of information and help the information flow between feature channels. After that, the output features are element-wise summed to obtain the feature ${F_1}^{S-C}$ with enriched channel information. The entire process is shown in Eq. (10). Compared to [28], in this paper, we perform a GAP operation on the feature ${F_1}^{S-C}$ to obtain the global weight for each feature channel. Let the input feature ${F_1}^{S-C}$ be denoted as $T \in R^{C\times H\times W}$, and apply the GAP operation to aggregate T into $\tilde{T} \in R^{C\times 1\times 1}$, as shown in Eq. (11). In Eq. (11), $m_{i,j}$ represents the value of the input feature T at point (i, j). After performing the GAP operation on feature ${F_1}^{S-C}$ to obtain the global weight for each feature channel, a convolutional layer is used to form a bottleneck structure, reducing the network's parameters from C^3/r to $C + C/r$. The obtained weights are then used for the Hadamard product with the features to obtain the new feature: ${F_1}^{S-C} = {F_1}^{S-C} \odot W_C$, with W_C representing the channel-wise weights obtained similarly to Eq. (4).

$$F_1^{S-C} = GWC(F_1^S) \oplus PWC(F_1^S), \left(GWC \in R^{\frac{\lambda C}{g}\times k\times k},\ PWC \in R^{(1-\lambda)C\times 1\times 1}\right) \tag{10}$$

$$\tilde{T}_k = \frac{1}{W \times H} \sum_{i}^{H} \sum_{j}^{W} m_{i,j} (1 \le i \le H, 1 \le j \le W) \tag{11}$$

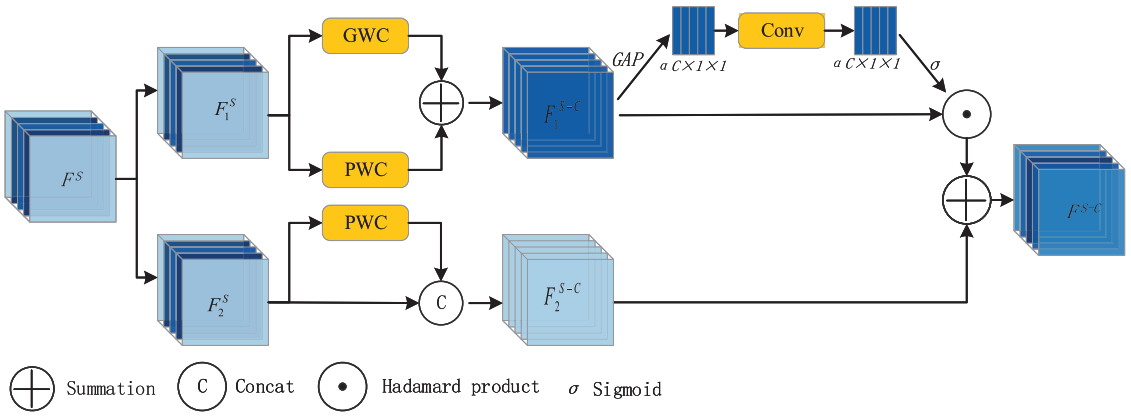

Figure 5: Channel information reconstruction

Since F_1^{S-C} contains fewer channel information and has a lot of redundant information, PWC is applied to extract features from it as a supplement to feature F_1^S. The features after PWC and feature F_2^S are concatenated along the channel dimension to obtain feature F_1^{S-C}. The entire process is shown in Eq. (11). After that, feature F_1^S is element-wise added with feature F_2^S to obtain the channel information-reconstructed feature F^{S-C}. The entire process is shown in Eq. (12).

$$F_2^{S-C} = Cat[F_2^S, PWC(F_2^S)], (PWC \in R^{(1-\xi)C\times 1\times 1}) \tag{12}$$

$$F^{S-C} = {F_1}^{S-C} \oplus {F_2}^{S-C} \tag{13}$$

The input image is processed through the feature extraction network ResNeXt-S to obtain the feature $\{F_2, F_3, F_4, F_5\}$ with an output channel of $\{256, 512, 1024, 2048\}$. Since higher-level features contain more semantic information, while lower-level features lack sufficient semantic information, it is necessary to fully extract the feature information of small objects. To achieve this, the designed FFBCA is first used for feature fusion. The feature F_5 is upsampling by a factor of 2 to match the spatial size of feature F_4, and then the channels are concatenated. Subsequently, the new feature map is reconstructed through spatial-channel information reconfiguration to obtain the feature $M_4(14 \times 14 \times 1024)$. The entire process is shown in Eq. (14), where $S^{2\uparrow}$ denotes the 2x upsampling; *SIR* and *CIR* represent spatial information reconstruction and channel information reconstruction, respectively. Similarly, the calculation process for the feature $M_3(28 \times 28 \times 512)$ and $M_4(56 \times 56 \times 256)$ is shown in Eqs. (15) and (16).

$$M_4^{14\times14\times1024} = CIR(SIR\{Cat[F_4^{14\times14\times1024}, S^{2\uparrow}(M_5^{7\times7\times2048})]\}) \tag{14}$$

$$M_3^{28\times28\times512} = CIR(SIR\{Cat[F_3^{28\times28\times512}, S^{2\uparrow}(M_4^{14\times14\times1024})]\}) \tag{15}$$

$$M_2^{56\times56\times256} = CIR(SIR\{Cat[F_2^{56\times56\times256}, S^{2\uparrow}(M_3^{28\times28\times512})]\}) \tag{16}$$

2.3 Skip Residual Connection Fusion Module

The top-down layer-wise fusion method may lead to the loss of some critical semantic information in high-level features, which is especially detrimental to the detection of small objects. To address this issue, this paper designs a Skip Residual Connection Fusion (SRCF) Module. SRCF uses a skip connection approach to directly combine low-level detail information with high-level semantic information, thereby avoiding information loss during transmission and enhancing the semantic representation in the feature map. Fig. 6a shows the basic residual structure, where original feature map information is introduced into the output, allowing the network to learn changes relative to the original feature map, further reducing the information loss in feature extraction. Fig. 6b shows the SRCF designed in this paper, where the concept of residual connection is introduced during upsampling feature fusion.

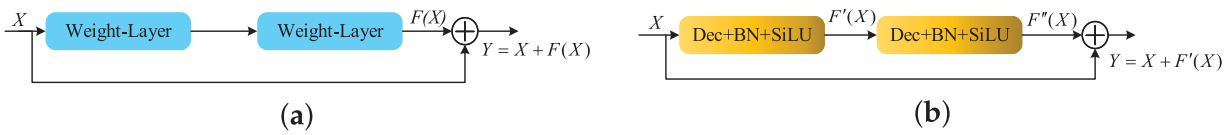

Figure 6: (**a**): Schematic diagram of residual structure. (**b**): Skip residual connection fusion

SRCF first applies two Deconvolution (kernel size: 1 × 1 and stride 2), Batch Normalization, and the adaptive activation function SiLU to upsampling the input high-level feature map to the spatial size of the lower-level feature map. Assume the input feature is $X \in R^{C\times H\times W}$, which is first upsampling to $F'(X) = \mathcal{S}(\mathcal{BN}[\mathcal{D}ec(X)])$. After the second upsampling, it is fused with the lower-level feature map, as shown

in the Eq. (17). In Eq. (17), $\mathcal{D}ecF'(X)$ represents deconvolution; $\mathcal{BN}$ represents Batch Normalization; $\mathcal{S}$ represents the SiLU activation function; and $\oplus$ represents element-wise addition. When performing feature fusion with SRCF, two upsampling fusion strategies for F_4 and F_2 are designed, namely AF and BF as shown in Fig. 1. AF represents upsampling F_4 before feature fusion and then fusing with F_2. BF represents performing feature fusion on F_4 first, followed by upsampling and fusion with F_2. In this paper, the BF approach is used for upsampling and feature fusion.

$$Y = \mathcal{S}(\mathcal{BN}\{\mathcal{D}ec[F'(X)]\}) \oplus X \ (X \in {}^{C\times H\times W}) \tag{17}$$

3 Experimental Results and Analysis

3.1 Experimental Setup

The software and hardware environment configuration used in the experiments in this paper is shown in Table 1. During training, the Batch Size is set to 2, the number of epochs is set to 25, the SGD optimizer is used, the initial learning rate is 0.01, the momentum is 0.9, and the cosine annealing algorithm is used to decrease the learning rate, with a minimum learning rate of 0.0001.

Table 1: Experimental setup

Configuration	Information
Experimental platform	Ubuntu 20.04.3
GPU model	NVIDIA RTX3090 24 GB
GPU acceleration	CUDA 11.3.0
Programming language	Python3.8
Deep learning framework	PyTorch1.12.0

3.2 Evaluation Metrics

In order to accurately evaluate the performance of the improved model in this paper, Average Precision (AP) and mean Average Precision (mAP) are used as evaluation metrics. AP represents the area under the curve plotted by Precision (P) and Recall (R) against the coordinate axes. The expression for AP is given in Eq. (19), where the calculation formulas for P and R are shown in Eq. (18). TP, FP, and FN represent true positives, false positives, and false negatives, respectively. mAP represents the mean of the AP for each class in the dataset, and its expression is given in Eq. (19).

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}. \tag{18}$$

$$AP = \int P(R)dR, mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{19}$$

3.3 Experimental Data and Results

3.3.1 Ochotona Curzoniae Object Detection

The Ochotona Curzoniae dataset used in this study was collected by our research team from the Gannan grassland in the northeastern Tibetan Plateau (longitude 101°35′36″–102°58′15″, latitude 33°58′21″–34°48′48″) and annotated according to the PASCAL VOC dataset format. The dataset contains a total of 1000 images, which are divided into training, validation, and test sets in a 6:2:2 ratio. The data

distribution in the Ochotona Curzoniae dataset is shown in Fig. 7. Fig. 7a shows the ratio of the height and width of the Ochotona Curzoniae targets to the height and width of the entire image. It can be observed that the ratio of the height and width of the Ochotona Curzoniae targets is mostly less than 20% compared to the entire image. Fig. 7b shows the histogram of the ratio of the area occupied by the pixels of Ochotona Curzoniae targets to the total image area in the dataset. From Fig. 7b, it can be seen that the pixel ratio of many Ochotona Curzoniae targets is less than 6%.

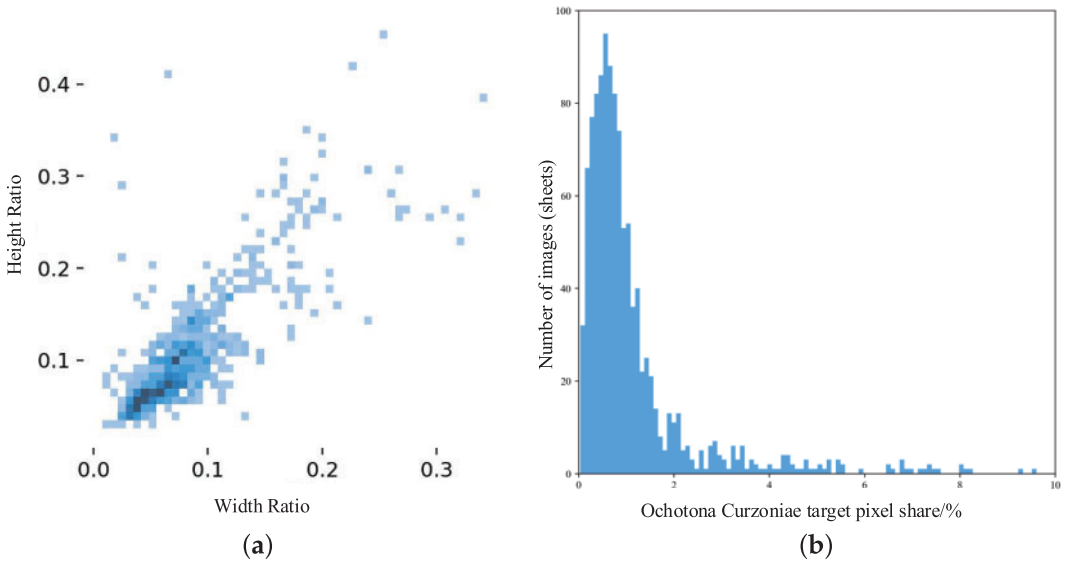

Figure 7: Analysis of the Ochotona Curzoniae dataset. (**a**) is distribution of Ochotona Curzoniae target height and width ratios; (**b**) is distribution of Ochotona Curzoniae target pixels as a percentage of area

Fig. 8 visualizes the attention of the original ResNeXt feature extraction network and the ResNeXt-S feature extraction network constructed in this study on different regions of the input image, with red indicating high-activation regions, i.e., the regions of interest. From Fig. 8b, it can be observed that the original ResNeXt does not pay enough attention to the Ochotona Curzoniae targets, and the attended regions are too large, which often leads to the waste of computational resources. Fig. 8c shows the visualization of the attention of the ResNeXt-S constructed in this study to the Ochotona Curzoniae targets. Compared to ResNeXt, ResNeXt-S pays more attention to the edges of the Ochotona Curzoniae targets, and the regions of interest are more concentrated. This is because the SimAM introduced in this study enhances the network's ability to capture detailed features of the Ochotona Curzoniae, leading to more accurate recognition of the Ochotona Curzoniae targets.

Table 2 shows the AP values of FSSD, TDFSSD, FPN, FFBAM, Faster R-CNN, SSD, and the object detection model designed in this study on the Ochotona Curzoniae dataset. FSSD512, TDFSSD512, and SSD512 indicate that the input image size for the models is 512 × 512 pixels. The detection AP value of the model proposed in this study for Ochotona Curzoniae is 92.3%, which is higher than FSSD512 (84.6%), TDFSSD512 (81.3%), FPN (86.5%), FFBAM (88.5%), Faster R-CNN (89.6%), and SSD512 (88.6%) by 7.7%, 11.0%, 5.8%, 3.8%, 2.7%, and 3.7%, respectively. This indicates that the model proposed in this study has higher detection accuracy than the models in the aforementioned literature.

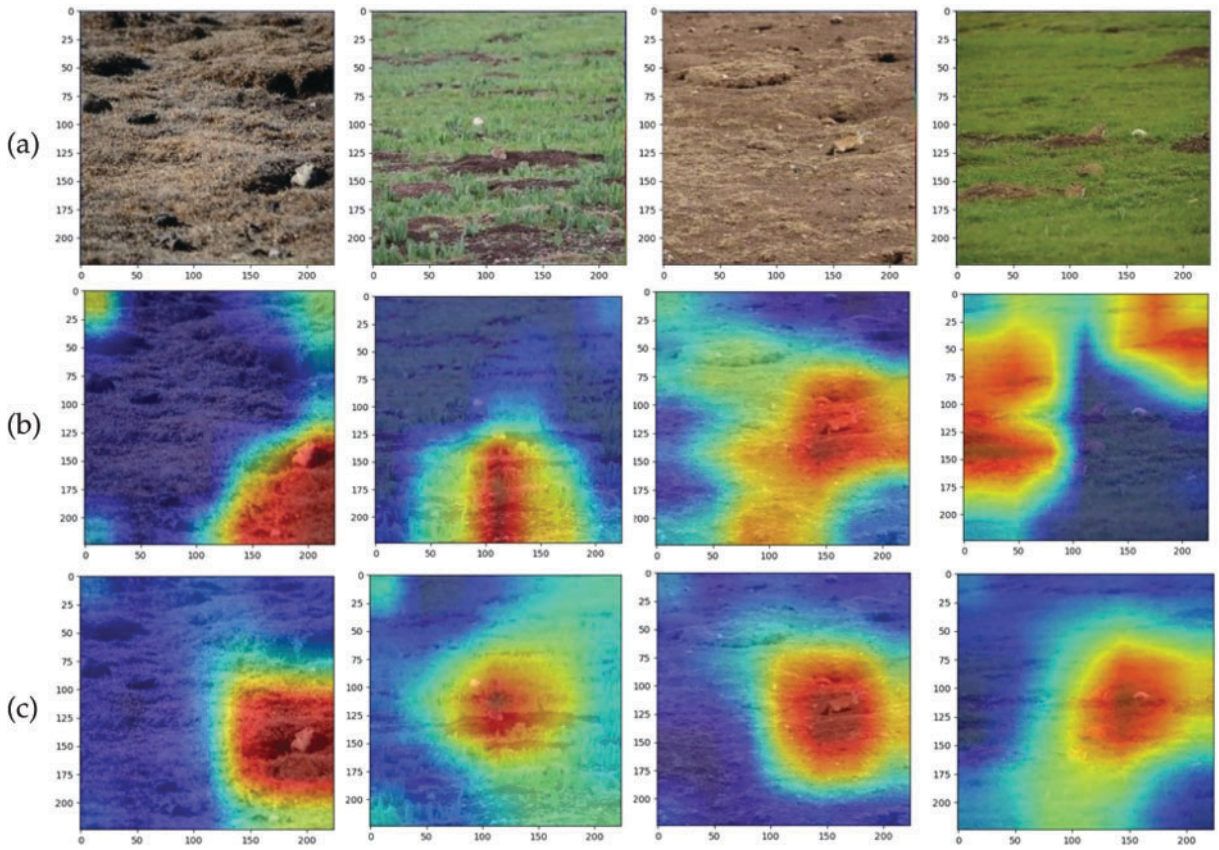

Figure 8: Comparison of class heat maps for different feature extraction networks. (**a**) Original Image; (**b**) ResNeXt; (**c**) ResNeXt-S

Table 2: AP values of object detection in the Ochotona Curzoniae dataset for different models

Model	FSSD512	TDFSSD512	FPN	FFBAM	Faster R-CNN	SSD512	Ours
AP (%)	84.6	81.3	86.5	88.5	89.6	88.6	92.3

The visualized object detection results from the randomly selected Ochotona Curzoniae test set are shown in Fig. 9. From Fig. 9a, it can be seen that the Ochotona Curzoniae images have small-sized targets with low pixel ratio and camouflage coloration that blends with the surrounding environment and are obscured by vegetation, making the features less distinct. From Fig. 9b, it can be seen that Faster R-CNN suffers from missed detections and false positives when detecting Ochotona Curzoniae targets. This is because the backbone network of Faster R-CNN undergoes multiple pooling operations, leading to the loss of feature information, making it difficult for the network to capture the key feature information of Ochotona Curzoniae. Additionally, Faster R-CNN only performs regression and classification at the high levels, neglecting the fine-grained information in the low-level features. This study improves on Faster R-CNN by constructing two structures: FFBCA and SRCF. FFBCA fuses high-level and low-level features, enabling the feature map to contain both detail information beneficial for localization and semantic information helpful for classification, while the attention mechanism suppresses redundant information generated during feature fusion, allowing the network to focus on more fine-grained features. SRCF introduces high-level semantic

information into low-level features, preventing the loss of semantic information when the model identifies Ochotona Curzoniae targets.

Figure 9: Example of qualitative detection in an Ochotona Curzoniae test set-up. (**a**) Original Image; (**b**) Faster R-CNN; (**c**) Ours

3.3.2 Object Detection on the NWPU VHR-10 Dataset

The NWPU VHR-10 is a high-resolution remote sensing image dataset released by Northwestern Polytechnical University in 2014. It contains 10 categories: airplanes, ships, oil tanks, baseball fields, tennis courts, basketball courts, athletic fields, ports, bridges, and cars. The dataset includes 650 images with targets and 180 background images. The dataset is divided into training, validation, and test sets in a 2:2:6 ratio.

To analyze the effectiveness of the proposed model, its performance is compared with several different object detection models on the NWPU VHR-10 dataset. The comparison results are shown in Table 3. From Table 3, it can be seen that the proposed model achieves an mAP of 91.1%, which is higher by 13.8%, 14.0%, 22.2%, 8.5%, 7.4%, 4.9%, and 3.0% compared to FSSD512 (77.3%), TDFSSD512 (77.1%), FPN (68.9%), FFBAM (82.6%), YOLOV5s (83.7%), Faster R-CNN (86.2%), and SSD512 (88.1%), respectively. This indicates that the proposed model has higher detection accuracy than the models in the mentioned references.

Table 3: Comparison of detection results for different models on the NWPU VHR-10 dataset

Model	Backbone	f_{mAP} (%)	f_{AP}@0.5 (%)									
			Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle
FSSD 512	VGG16	77.3	90.9	86.3	69.3	89.7	78.2	75.6	90.6	63.8	59.1	69.3
TDFSSD 512	VGG16	77.1	90.8	85.4	74.9	89.9	77.8	72.3	84.6	67.3	51.0	76.8
FPN	ResNet 101	68.9	90.8	82.8	76.7	89.7	73.9	61.4	66.5	64.7	15.0	67.7
FFBAM	ResNet 101	82.5	97.1	89.3	88.2	90.1	79.4	72.9	90.5	81,3	57,7	78,7
YOLOV5s	CSP Darknet53	83,7	97,4	77,1	95,2	95.9	81.2	75.3	98.0	88.5	62.3	66.5
Faster R-CNN	ResNet 101	86.2	90.8	89.4	89.5	90.1	87.4	84.3	90.8	81.4	75.0	83.2
SSD512	VGG16	88.1	99.3	80.5	98.0	86.7	81.4	89.3	79.4	91	86,8	88,9
Ours	ResNeXt50-Sim	91.1	97.9	87.0	96.3	96.2	95.5	87.5	96.2	86.2	78.5	89.5

Fig. 10 shows the qualitative detection examples randomly selected from the NWPU VHR-10 test set. From Fig. 10b, it can be observed that Faster R-CNN causes a large number of missed detections for targets such as oil tanks, which are dense, small, and have similar color to the background, and it also causes false positives for bridges. In contrast to the original model, the proposed model does not exhibit the above issues. This is because the FFBCA constructed in this study, after feature fusion, suppresses spatial and channel redundancy information through the attention mechanism, optimizing feature extraction, allowing the network to capture more detailed information of small or dense targets in the image. At the same time, SRCF combines features from different layers, allowing the network to leverage deep semantic information, thereby improving the model's performance in handling small targets.

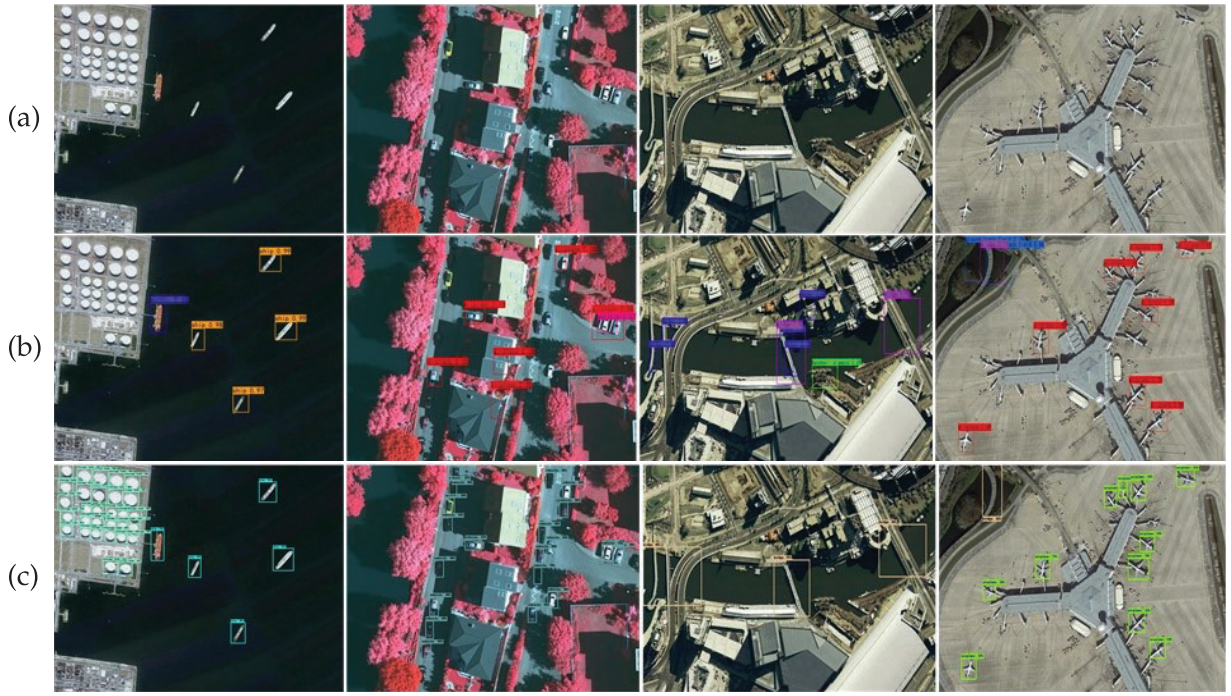

Figure 10: NWPU VHR-10 test set qualitative detection example. (**a**) Original image; (**b**) Faster R-CNN; (**c**) Ours

4 Discussion

This section presents the ablation experiments on various modules designed in the proposed model based on the experiments above. It aims to determine the impact of each module on the model's accuracy and its significance, thereby providing a more reasonable explanation for the constructed model. The ablation experiments were conducted on the NWPU VHR-10 dataset.

4.1 Ablation Experiments

To further analyze the impact of different modules on the accuracy of the Ochotona Curzoniae target detection model and verify the effectiveness of the proposed modules, several ablation experiments were conducted. The results are shown in Table 4, where '✓' represents the addition of the module and '–' represents its absence. In these experiments, specific modules were added or removed to compare the impact of different modules on the accuracy of Ochotona Curzoniae target detection, thus evaluating the importance of each module and its contribution to the overall detection accuracy of the model.

Table 4: Result of ablation experiments

CB_X	FFBCA	SRCF	f_{mAP} (%)
–	–	–	84.8
✓	–	–	86.9
✓	–	✓	90.5
✓	✓	–	90.6
✓	✓	✓	91.1

In Table 4, it can be seen that the proposed CB_X module improves detection accuracy by 5.2% compared to the original ResNeXt (84.8%) backbone network. The proposed channel-space information reconstruction improves the detection accuracy by 0.6% compared to the lack of information reconstruction (90.0%), indicating that the proposed FFBCA can enhance the model accuracy. The proposed skip residual connection fusion method improves detection accuracy by 0.5% compared to the 'direct fusion' method (90.5%).

4.2 Impact of Sim-ResNeXt on Model Detection Accuracy

Table 5 shows the impact of different constructions of Sim-CB on the model detection accuracy. In the table, Model a, Model b, and Model c correspond to the three connection methods shown in Fig. 2. From the data in Table 5, it can be observed that using Sim Attention before convolution results in the most significant improvement in the model accuracy. This is because Sim Attention allocates a weight to each neuron based on its importance, thereby enhancing the feature map's representational ability. This enhanced feature map is then fed into the convolutional layer, allowing it to focus more on important features, further improving the network's performance.

Table 5: Impact of Sim-CB on model detection accuracy

	f_{mAP} (%)	f_{AP}@0.5 (%)									
		Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle
Model-a	88.8	95.9	85.6	95.2	94.8	93.8	87.6	95.9	80.5	71.4	87.6
Model-b	90.0	96.9	81.2	96.3	92.9	94.7	84.8	95.2	85.6	83.4	89.3
Model-c	86.9	94.9	80.4	94.9	96.4	94.8	3.4	92	79.8	75.3	76.6

4.3 Impact of FFBCA on Model Detection Accuracy

Table 6 shows the impact of spatial and channel information reconstruction on model detection accuracy. In the table, 'S' and 'C' represent the reconstruction of spatial and channel information reconstruction, respectively; 'S-C' and 'C-S' represent first spatial and then channel, or first channel and then spatial; 'S&C' (or 'C&S') means performing spatial and channel reconstruction separately and then combining the reconstructed features by adding elementwise. From Table 6, it can be seen that performing only channel or spatial information reconstruction does not significantly improve the model's accuracy and may even reduce it. This suggests that reducing redundancy in just one direction is insufficient and supports the rationality of spatial-channel information reconstruction. Based on experimental comparison, it was found that the spatial-channel (S-C) reconstruction yields the most significant improvement in model accuracy.

Table 6: Impact of FFBCA on model detection accuracy

	f_{mAP} (%)	f_{AP}@0.5 (%)									
		Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle
S	86.6	88.8	83.1	89.3	96.1	89.6	86.4	93.6	80	74.7	84.3
C	88.1	96.9	81.8	88.2	95.5	92.2	79.7	93.1	88.1	77.1	88.4
S-C	89.6	98.7	85.0	95.3	93.1	95.5	86.2	95.3	85.0	72.1	90.0
C-S	89.2	97.0	84.0	94.9	95.9	93.1	85.9	97.8	79.3	73.2	90.4
S&C (C&S)	88.7	98.0	81.0	95.6	94.5	95.3	88.1	95.8	88.1	62.3	88.7

4.4 Impact of SRCF on Model Detection Accuracy

Table 7 shows the impact of SRCF on model detection accuracy. It can be seen that when constructing the feature pyramid structure, introducing higher-level information into the lower-level features significantly improves the model's accuracy compared to the traditional top-down construction method. This is because the top-down feature fusion method, during the process of upsampling layer by layer, causes loss of feature information, which is particularly detrimental to small target detection. It is worth noting that regardless of whether AF or BF fusion paths are used, the mAP is improved by 3.8% and 4.3%, respectively, compared to not using the Skip Residual Connection Fusion. Therefore, the BF fusion path was adopted in this paper.

Table 7: Impact of SRCF on model detection accuracy

	f_{mAP} (%)	f_{AP}@0.5 (%)									
		Airplane	Ship	Storage tank	Baseball diamond	Tennis court	Basketball court	Ground track field	Harbor	Bridge	Vehicle
AF	90.0	98.3	85.5	97.0	93.7	86.4	91.3	84.4	90.0	85.8	87.9
BF	90.5	97.7	80.2	94.6	94.5	95.3	88.1	95.8	88.1	82.3	88.7

4.5 Comprehensive Comparative Analysis of FPS and mAP of Different Models

Fig. 11 shows the comparative analysis of the combined consideration of FPS and mAP for the different models, where the horizontal coordinates represent the mAP values, the vertical coordinates represent the FPS, and the area of the circle represents the importance of their combined consideration of the two.

As can be seen in Fig. 11, the FPS of our model (19.7) is slightly lower than that of Faster R-CNN (20.4), which is due to the fact that in this paper, ResNeXt-S with more convolutional layers and more complexity is used in the extraction of feature information by the backbone network, and at the same time, the designed FFBCA and SRCF are added into the feature fusion, and these modules improve the performance of feature extraction and fusion but also increase the computational burden, but an mAP of our model (92.3%) is higher than that of Faster R-CNN (89.6%).

The FPS of our model is higher than the multi-scale feature fusion-based models FPN (7.7) and FFBAM (7.5), but the FPS of our model is lower than that of FSSD512 (100), TDFSSD512 (41.1), YOLOV5s (150.0), and SSD512 (40.0), which is due to the fact that the above models don't need to generate the target detection in candidate regions but directly predict the target's bounding box and category, but our model needs to generate candidate regions in the image to be detected first and then predict the target's bounding box and category for each candidate region, which increases the detection time of the model. However, taking all factors into account, our model still has a significant accuracy advantage over other models.

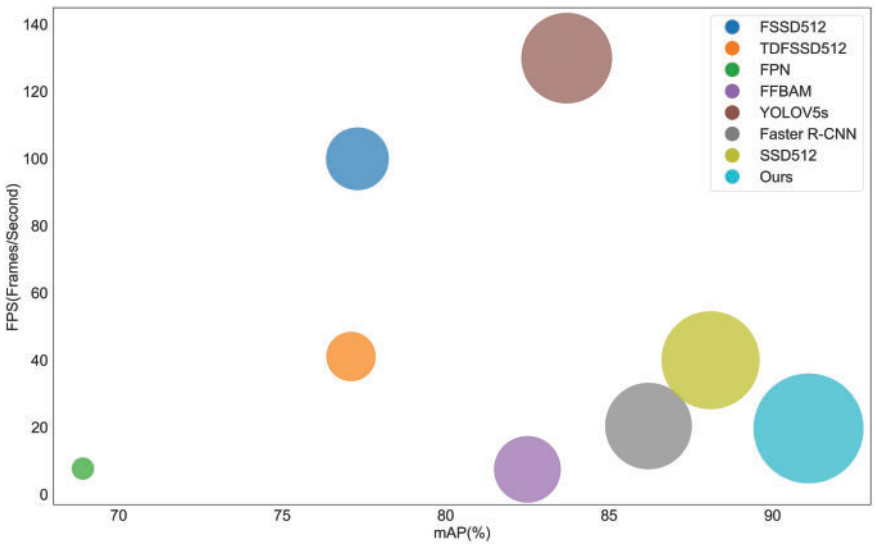

Figure 11: Comprehensive comparative of FPS and mAP of different models

5 Conclusion

The characteristics of Ochotona Curzoniae images in natural scenes, such as the small proportion of target pixels and unremarkable features, make few features available for target detection, and feature fusion can effectively enhance the feature representation of the target, but the information redundancy and high-frequency noise will be generated by the information repetition and semantic bias of the features at different levels in the top-down fusion process. In order to suppress the information redundancy and high-frequency noise generated during the feature fusion process, and to improve the accuracy of target detection for Ochotona Curzoniae, we designed a target detection model for Ochotona Curzoniae based on spatial-channel reconstruction convolutional attention mechanism feature fusion based on Faster R-CNN.

The first designed ResNeXt-S feature extraction network enhances the feature extraction capability for Ochotona Curzoniae targets by group convolution and attention mechanism. After that, the designed dual attention mechanism not only suppresses the redundant information in space but also suppresses the redundant information and high-frequency noise in the channel; finally, the designed jump residual link fusion reduces the loss of the high-level semantic information by introducing the high-level semantic information into the low-level, which enhances the low-level semantic information and reduces the loss of the high-level semantic information at the same time. The experimental results of target detection in the Ochotona Curzoniae dataset show that the target detection model proposed in this paper can effectively improve the accuracy of target detection for Ochotona Curzoniae.

Acknowledgement: The authors are grateful to all the editors and anonymous reviewers for their valuable comments and suggestions. We also extend our thanks to all the members who have contributed to this work with us.

Funding Statement: This document is the result of a research project funded by the National Natural Science Foundation of China (Grant Nos. 62161019, 62061024).

Author Contributions: The authors acknowledge their contributions as follows: Haiyan Chen provided the Ochotona Curzoniae image dataset and prepared the manuscript; Rong Li was responsible for the conception and design of the study, and the analysis and interpretation of the results. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data for this study comes from two main sources: 1. Data publicly available in the public repository: NWPU VHR-10. Data supporting the results of this study are publicly available in the Public Dataset at https://aistudio.baidu.com/datasetdetail/225242/0 (accessed on 18 June 2025). 2. The Ochotona Curzoniae dataset is not available due to confidentiality agreements. This dataset consists of actual images taken from the subject group and is a private dataset and it is not possible to share these images.

Ethics Approval: This study did not involve human or animal subjects, and therefore ethical approval was not applicable.

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

References

1. Lambert JP, Hartmann JV, Kun S, Riordan P. The effects of plateau pika *(Ochotona curzoniae)* presence and population control on the structure of an alpine grassland bird community. J Resour Ecol. 2023;14(1). doi:10.5814/j.issn.1674-764x.2023.01.005.
2. Wang X, Ye Z, Zhang C, Wei X. Effect of *Plateau pika* on soil microbial assembly process and co-occurrence patterns in the alpine meadow ecosystem. Microorganisms. 2024;12(6):1075. doi:10.3390/microorganisms12061075.
3. Su XX, Li XL, Sun HF, Song ZH, Li JX, Zhang J. Effects of *Plateau pika* and mowing disturbances on plant community and soil physical and chemical properties in alpine meadow. Acta Ecol Sin. 2024;44(22):10189–99. (In Chinese). doi:10.20103/j.stxb.202310272335.
4. Chen YY, Yang H, Bao GS, Pang XP, Guo ZG. Effect of the presence of *Plateau pikas* on the ecosystem services of alpine meadows. Biogeosciences. 2022;19(18):4521–32. doi:10.5194/bg-19-4521-2022.
5. Xu XT, Wang YM, Wang XZ, Li JN, Li J, Yang D, et al. Consequences of plateau pika disturbance on plant-soil carbon and nitrogen in alpine meadows. Front Plant Sci. 2024;15:1362125. doi:10.3389/fpls.2024.1362125.
6. Xu P, Zhang Y, Ji M, Guo S, Tang Z, Wang X, et al. Advanced intelligent monitoring technologies for animals: a survey. Neurocomputing. 2024;585:127640. doi:10.1016/j.neucom.2024.127640.
7. Lian HH, Zhao XL, Wang HC. Research on location monitoring of Plateau Pika using artificial intelligence object recognition technology. Qinghai Grassland. 2023;32(2):8–14.
8. Rozhnov VV, Salman AL, Yachmennikova AA, Lushchekina AA, Salman PA. Automated identification and counting of saigas (Saiga tatarica) by using deep convolutional neural networks in high-resolution satellite images. Biol Bull. 2024;51(5):1407–21. doi:10.1134/S1062359024608784.
9. Segura-Garcia J, Sturley S, Arevalillo-Herraez M, Alcaraz-Calero JM, Felici-Castell S, Navarro-Camba EA. 5G AI-IoT system for bird species monitoring and song classification. Sensors. 2024;24(11):3687. doi:10.3390/s24113687.
10. Zhang Z, Zhu W. YOLO-MFD: remote sensing image object detection with multi-scale fusion dynamic head. Comput Mater Contin. 2024;79(2):2547–63. doi:10.32604/cmc.2024.048755.
11. Wei W, Cheng Y, He J, Zhu X. A review of small object detection based on deep learning. Neural Comput Appl. 2024;36(12):6283–303. doi:10.1007/s00521-024-09422-6.
12. Rohan A, Rafaq MS, Hasan MJ, Asghar F, Bashir AK, Dottorini T. Application of deep learning for livestock behaviour recognition: a systematic literature review. Comput Electron Agric. 2024;224:109115. doi:10.1016/j.compag.2024.109115.
13. Zhang T, Zhuang Y, Wang G, Dong S, Chen H, Li L. Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery. IEEE Trans Geosci Remote Sens. 2021;60:5608720. doi:10.1109/TGRS.2021.3108476.
14. Yu W, Zhang J, Liu D, Xi Y, Wu Y. An effective and lightweight full-scale target detection network for UAV images based on deformable convolutions and multi-scale contextual feature optimization. Remote Sens. 2024;16(16):2944. doi:10.3390/rs16162944.
15. Li ZX, Zhou FQ. FSSD: feature fusion single shot multibox detector. arXiv:1712.00960. 2017.
16. Pan H, Jiang J, Chen G. TDFSSD: top-down feature fusion single shot MultiBox detector. Signal Process Image Commun. 2020;89:115987. doi:10.1016/j.image.2020.115987.

17. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. IEEE. p. 936–44. doi:10.1109/CVPR.2017.106.
18. Chen HY, Zhen XJ, Zhao TT. An attention mechanism feature fusion for small target detection model. J Huazhong Univ Sci Technol (Nat Sci Ed). 2023;51(3):60–6.
19. Lian J, Yin Y, Li L, Wang Z, Zhou Y. Small object detection in traffic scenes based on attention feature fusion. Sens. 2021;21(9):3031. doi:10.3390/s21093031.
20. Guo C, Fan B, Zhang Q, Xiang S, Pan C. AugFPN: improving multi-scale feature learning for object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 13-19, 2020; Seattle, WA, USA. IEEE; 2020. p. 12592–601. doi:10.1109/cvpr42600.2020.01261.
21. Zou F, Xiao W, Ji W, He K, Yang Z, Song J, et al. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. Neural Comput Appl. 2020;32(18):14549–62. doi:10.1007/s00521-020-04893-9.
22. Zhao X, Chen Y, Guo J, Zhao D. A spatial-temporal attention model for human trajectory prediction. IEEE/CAA J Autom Sinica. 2020;7(4):965–74. doi:10.1109/jas.2020.1003228.
23. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell. 2020;42(8):2011–23. doi:10.1109/tpami.2019.2913372.
24. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Computer Vision-ECCV 2018. Cham: Springer International Publishing; 2018. p. 3–19. doi:10.1007/978-3-030-01234-2_1.
25. Ren J, Zhang Z, Fan J, Zhang H, Xu M, Wang M. Robust low-rank deep feature recovery in CNNs: toward low information loss and fast convergence. In: 2021 IEEE International Conference on Data Mining (ICDM); 2021 Dec 7–10; Auckland, New Zealand; IEEE; 2021. p. 529–38. doi:10.1109/ICDM51629.2021.00064.
26. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
27. Wang L, Wang L, Wang Q, Bruzzone L. RSCNet: a residual self-calibrated network for hyperspectral image change detection. IEEE Trans Geosci Remote Sens. 2022;60:5529917. doi:10.1109/TGRS.2022.3177478.
28. Li J, Wen Y, He L. SCConv: spatial and channel reconstruction convolution for feature redundancy. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 17–22, 2023; Vancouver, BC, Canada: IEEE. p. 6153–62. doi:10.1109/CVPR52729.2023.00596.