scientific reports

OPEN

Segmentation of plateau zokor mounds in alpine meadows from UAV images using an improved UNet network

Yang Yang[1], Lianguo Wang[1]✉ & Limin Hua[2]✉

Plateau zokor mounds, created by the burrowing activity of Plateau zokor, cause significant damage to crops, grasslands, and infrastructure, particularly in the alpine meadows of the Tibetan Plateau. Traditional field surveys are inefficient and labor-intensive, limiting the ability to conduct large-scale monitoring. Accurate detection of zokor mounds is essential for effective rodent control and sustainable grassland management. This study introduces VGG–Dice–PSA UNet(VDP_UNet), an enhanced deep learning model designed to segment zokor mounds from UAV imagery captured at 30 m. Based on the UNet architecture, VGG16 is used to replace the original UNet backbone, enabling the model to capture global contextual information and enhance feature extraction in complex backgrounds. Additionally, a Polarized Self-Attention (PSA) module is integrated into the feature fusion stage following the encoder–decoder skip connections to better capture fine-grained semantic features related to zokor mounds. To reduce overfitting and address class imbalance, Dice Loss is introduced during training. VDP_UNet was trained and evaluated on a custom high-resolution zokor mound dataset. It achieved an IoU of 51.99%, MIoU of 75.63%, mean Pixel Accuracy of 82.66%, Precision of 71.44%, FPS of 42.13 f/s, Accuracy of 99.27%, and an F1-score of 68.41%, outperforming recent deep learning models. Experimental results indicate that the proposed VDP_UNet model efficiently segments zokor mounds in alpine meadows, markedly improving the extraction of mound features from UAV images. Furthermore, this study establishes a practical foundation for estimating mound areas in real sample plots and provides solid technical support for rodent control and the sustainable development of alpine ecosystems.

Keywords Semantic segmentation, VDP_UNet network, Alpine meadow, Zokor mound dataset, Rodent management

The plateau zokor (*Eospalax baileyi*), a member of the order Rodentia and family Spalacidae[1], is a subterranean rodent species endemic to the grasslands of the Tibetan Plateau. Its burrowing and mound-building activities strongly influence primary productivity and herbivore interactions, while accelerating soil nutrient turnover, enhancing microbial activity[2], and promoting decomposition processes[3]. The presence of zokor mounds not only disrupts plant community succession and affects carbon sequestration[4], but also reduces available grazing area[5], accelerates soil erosion[6], and contributes to a decline in biodiversity[7], ultimately diminishing ground cover and productivity in alpine meadows[8]. These impacts have made zokor activity one of the key drivers of grassland degradation[9]. Furthermore, the spread of zokor mounds poses serious threats to forestry operations and grassland ecological security[10–12]. Therefore, the precise identification and extraction of zokor mounds from UAV images is essential for enhancing the effectiveness of rodent damage control efforts in alpine grasslands and supporting long-term ecological sustainability.

Currently, due to limited monitoring capabilities, rodent control in alpine meadows largely relies on indiscriminate large-scale extermination efforts. Although this approach can suppress rodent outbreaks in the short term, it overlooks the multiple ecological roles of the plateau zokor within grassland ecosystems and

[1]College of Information Science and Technology, Gansu Agricultural University, Lanzhou 730070, China. [2]College of Grassland Science, Key Laboratory of Grassland Ecosystem of the Ministry of Education, Engineering and Technology Research Center for Alpine Rodent Pest Control, Gansu Agricultural University, National Forestry and Grassland Administration, Lanzhou 730070, China. ✉email: lianguoWang@163.com; hualm@gsau.edu.cn

compromises both ecological stability and functional integrity. Therefore, scientifically monitoring rodent activity is a fundamental prerequisite for implementing tiered management strategies that balance biodiversity conservation with effective control measures. At present, rodent damage monitoring in China's grasslands primarily relies on manual field surveys[13], which focus on counting burrows, bare patches, and mounds[14–16], while often neglecting the importance of affected area. Traditional monitoring methods are inefficient and costly, making it difficult to simultaneously extract the number and area of zokor mounds over large regions[17] or meet the demands of high-precision monitoring[18]. The emergence of remote sensing technologies offers new opportunities for rodent damage assessment. For example, Wang et al.[19] utilized UAV imagery combined with supervised classification methods to investigate zokor mound distribution across plots of varying densities in Ruoergai County, Sichuan Province, enabling the classification of rodent damage severity and risk prediction. In Maqu County, Gansu Province, Hua et al.[20] employed UAV remote sensing and hierarchical sampling to establish a three-level sample system and estimate the area affected by plateau pika damage. Although these studies represent progress in remote sensing-assisted monitoring of rodent mounds, the process still heavily depends on manual interpretation, with algorithmic models playing only a supplementary role[21]. Moreover, supervised classification performs poorly in complex scenarios, limiting its applicability. Therefore, there is an urgent need to develop rodent mound monitoring approaches that integrate remote sensing imagery with deep learning techniques to improve overall monitoring efficiency and application scope.

UAV imaging enables rapid acquisition of large-scale, high-resolution remote sensing data, providing a strong foundation for fine-grained target identification. When combined with machine learning, it allows efficient detection of zokor mound locations and contours, greatly reducing the need for manual image analysis[22]. As a result, the integration of UAV remote sensing imagery and machine learning represents a vital approach and emerging research direction for efficient zokor mound monitoring. In existing studies, Qi et al.[23] proposed a detection framework that combines UAV images, object-based image analysis, and the correlation-based feature selection algorithm to efficiently extract rodent burrow patches in desert grasslands. However, the study was limited by a small number of samples. Li et al.[24] applied "3S" technologies in conjunction with the maximum likelihood method and decision tree classification to estimate the affected area of rodent damage in the Altun Mountains. Yet, the image resolution was insufficient for small-scale or topographically complex regions. Sandino et al.[25] developed an automated method integrating UAV-based hyperspectral imaging, machine learning, and image processing to detect termite mounds, though the accuracy of the model still requires improvement. In recent years, the advancement of deep learning has greatly enhanced the efficiency of image processing, especially in feature extraction, and has driven the intelligent analysis of remote sensing data[26]. In the context of rodent monitoring, deep learning has been widely used for tasks such as burrow detection[27,28], population estimation[29], and predictive modeling. However, studies focusing on extracting zokor mound areas using deep learning methods remain relatively scarce. Given that deep learning–based semantic segmentation algorithms can automatically learn complex image features and demonstrate superior accuracy and generalization performance[30], developing a precise deep learning model for the automated detection of zokor mounds across large-scale alpine meadow regions offers a more efficient and intelligent solution for rodent damage monitoring.

To address the aforementioned challenges, this study proposes a deep learning-based segmentation model named VGG–Dice–PSA UNet(VDP_UNet). The model enhances the traditional UNet architecture by replacing its encoder with VGG16, enabling improved capture of global contextual information from input images. To further strengthen feature representation under complex background conditions, a Polarized Self-Attention (PSA) module is integrated into the feature fusion stage following the encoder–decoder skip connections. Additionally, Dice Loss is adopted to alleviate class imbalance between zokor mound and background samples. We conducted a comprehensive performance comparison between VDP_UNet and both traditional and state-of-the-art segmentation methods. The results demonstrate that VDP_UNet consistently outperforms other approaches, accurately extracting zokor mounds in challenging environments and exhibiting strong generalization capabilities. This method offers robust technical support for the efficient monitoring of zokor mounds in alpine meadows.

Materials and methods

Overview of the study area

The study area is located in Zhuaxixiulong Town, Tianzhu Tibetan Autonomous County, Gansu Province (37°12′13″N, 102°46′11″E; elevation: 2,890.38 m), adjacent to the eastern edge of the Tibetan Plateau (As shown in Fig. 1). It represents a typical habitat for the plateau zokor. The region is characterized by a plateau continental climate, marked by significant annual temperature variation and pronounced diurnal temperature differences. The soil type is subalpine meadow soil, and the vegetation consists of typical alpine meadow. The climate features two primary seasons: a warm season from May to October and a cold season from November to April of the following year. The area experiences no absolute frost-free period, with an average annual temperature of approximately − 0.1 °C. The mean annual precipitation is 416.0 mm, most of which occurs between July and September[31]. The plateau zokor is the only subterranean rodent species in this region[32]. Its burrowing activities result in the formation of mounds on the soil surface[33], producing a characteristic "mound–vegetation" mosaic pattern across the grassland landscape.

Data collection and preprocessing

Data acquisition

In this study, a DJI Mavic 2 Pro quadcopter UAV (https://www.dji.com/mavic-2) equipped with a Hasselblad L1D-20c camera was used to capture RGB images of plateau zokor mounds. The camera features a 1-inch, 20-megapixel CMOS image sensor, enabling the acquisition of high-resolution images. Data collection was conducted on April 12, 13, and 15, 2021, as well as on June 5, 13, 15, 17, and 19, 2023, covering both the returning

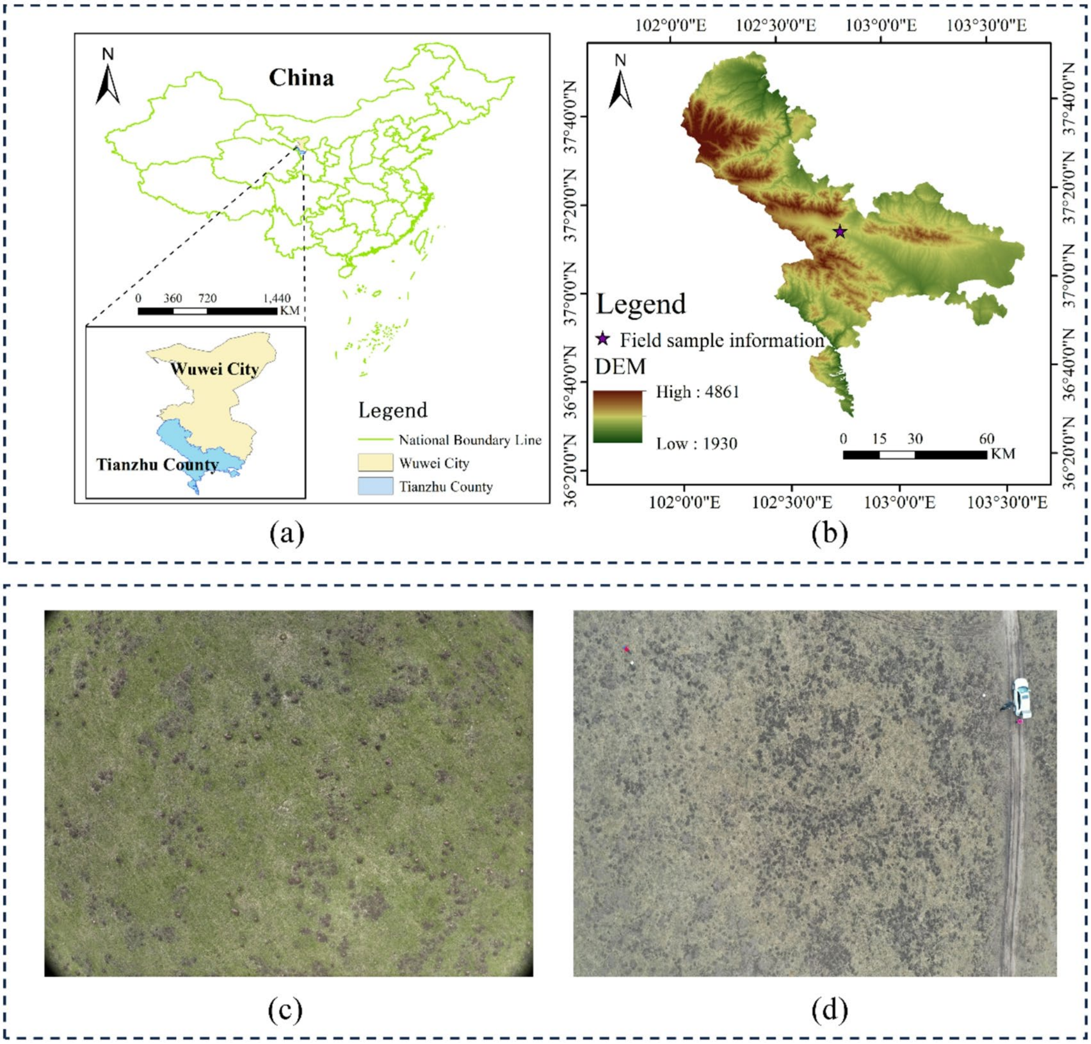

Fig. 1. Overview map of the study area. (**a**) Administrative divisions of Tianzhu county. (**b**) Digital elevation model of Tianzhu county. (**c**) Representative UAV image of plateau zokor mounds during the peak grass period. (**d**) Representative UAV image of plateau zokor mounds during the returning green period.

green period and peak grass period of the alpine meadow vegetation cycle. The UAV followed pre-programmed flight paths at an altitude of 30 m, capturing multi-directional images with both forward and side overlap rates maintained at 75% to ensure full coverage. The resolution of the collected zokor mound images was 5280 × 3956 pixels. To ensure image quality and minimize interference from environmental factors such as lighting and wind, all flights were conducted under clear skies with calm wind conditions and ample sunlight. In total, 876 high-quality zokor mound images were acquired for further analysis.

Data preprocessing

Initial annotations were generated using the automatic labeling tool on the Roboflow platform (Roboflow, Des Moines, Iowa, US). Targets were classified into two categories: zokor mounds (labeled "ZM") and non-mound areas. To support the subsequent segmentation task, category and location data were saved in JSON format. Annotation was carried out using a tagging system powered by Grounding DINO, which identifies and labels surface mounds of varying sizes formed by zokor digging activity. Developed by IDEA Research, Grounding DINO is an open-vocabulary object detection model that integrates object recognition with multimodal understanding[34]. By leveraging natural language prompts, it enables fast and accurate identification of multiple relevant objects within an image. Unlike traditional detection models, Grounding DINO incorporates a language understanding module, giving it the ability to recognize unfamiliar categories in open-world scenarios[35]. In this study, it proved especially effective in handling densely distributed and complex zokor mound scenes, significantly improving labeling efficiency and accuracy—thereby offering strong support for dataset development.

Following the preliminary annotation using DINO, experts conducted a meticulous image-by-image review with Labelme to ensure the accurate labeling of plateau zokor mounds. Images exhibiting excessive overexposure or underexposure that compromised visual details were excluded. As a result, an initial dataset of 869 images was finalized. This dataset was then randomly split into a training set (90%) and a validation set (10%), with no separate test set used in this experiment. Because deep learning models are typically sensitive to sample size, five data augmentation techniques—translation, flipping, mirroring, noise addition, and brightness adjustment—were applied to the training images to prevent overfitting and improve model robustness. After augmentation, a total of 5,214 images were generated, forming the alpine meadow plateau zokor mound dataset, as shown in Fig. 2. Among these, 4,692 images were used for training, and 522 images were used for performance validation during training. For image annotation, OpenCV was used to create pixel-level labels, assigning a value of 1 to annotated regions (zokor mounds) and 0 to unannotated regions (non-mound areas), facilitating subsequent semantic segmentation training.

Network architecture

To mitigate the issues of gradient vanishing and exploding during deep neural network training, and to enhance the ability to extract zokor mound features from remote sensing imagery, we optimized the traditional UNet architecture and proposed an improved model specifically designed for zokor mound extraction in UAV images. Specifically, the original UNet encoder was replaced with the VGG16 backbone, which offers a favorable balance between depth and computational efficiency. This replacement enhances the model's capability to represent complex spatial structures and RGB features in remote sensing images, enabling more accurate identification of subtle differences in zokor mounds. In addition, to further strengthen the model's focus on key regions, we introduced the PSA module after the skip connections to enhance feature representation. This module improves the model's sensitivity to local spatial structure and semantic information, especially in scenes with complex backgrounds or densely distributed targets. Given that zokor mounds typically occupy a small portion of remote sensing images—resulting in a pronounced class imbalance—we adopted Dice Loss as the loss function. This choice improves the model's training performance under imbalanced positive and negative samples and enhances detection precision. Based on the above improvements, we propose the VDP_UNet segmentation model, architecture is shown in Fig. 3.

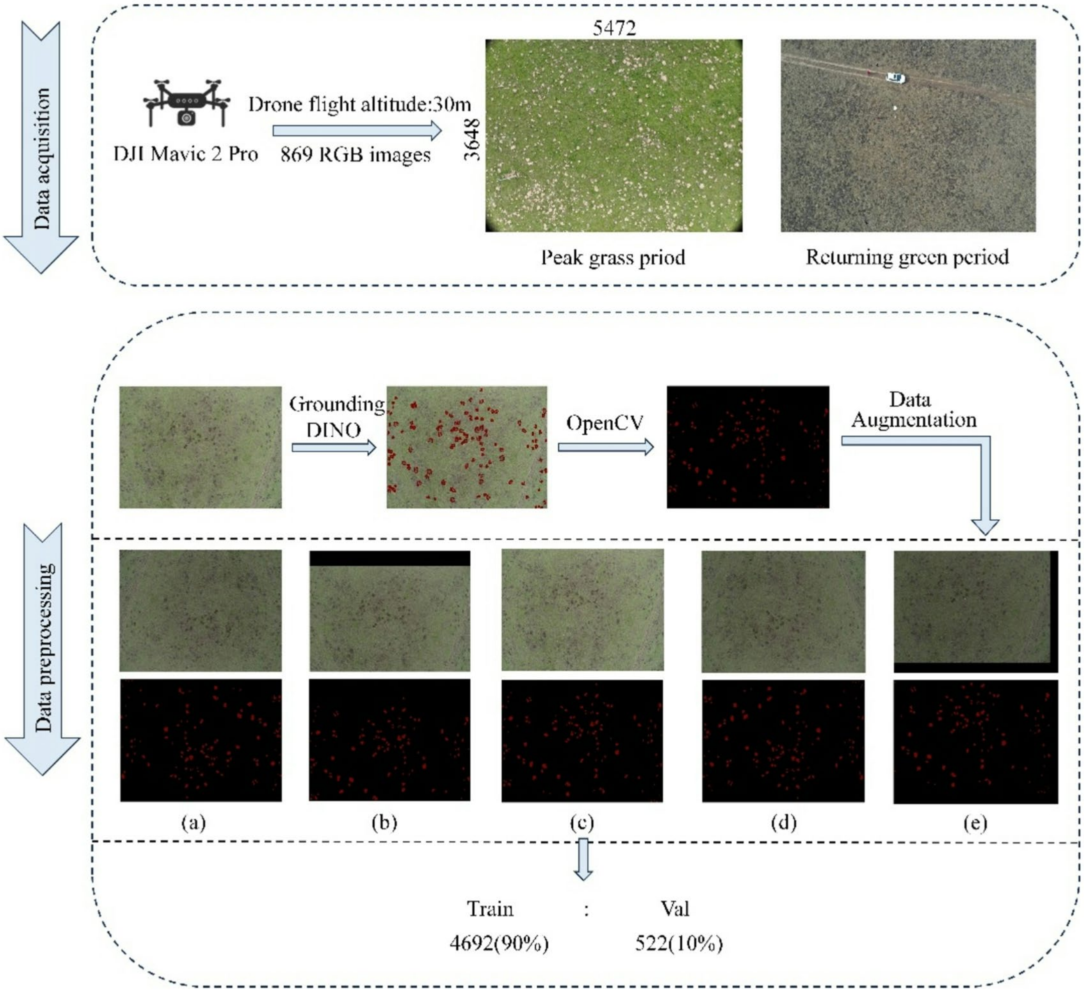

Fig. 2. Flowchart of data collection and processing: (**a**) flip, add noise, and adjust brightness; (**b**) translate and add noise; (**c**) add noise; (**d**) flip; (**e**) translate and adjust brightness.

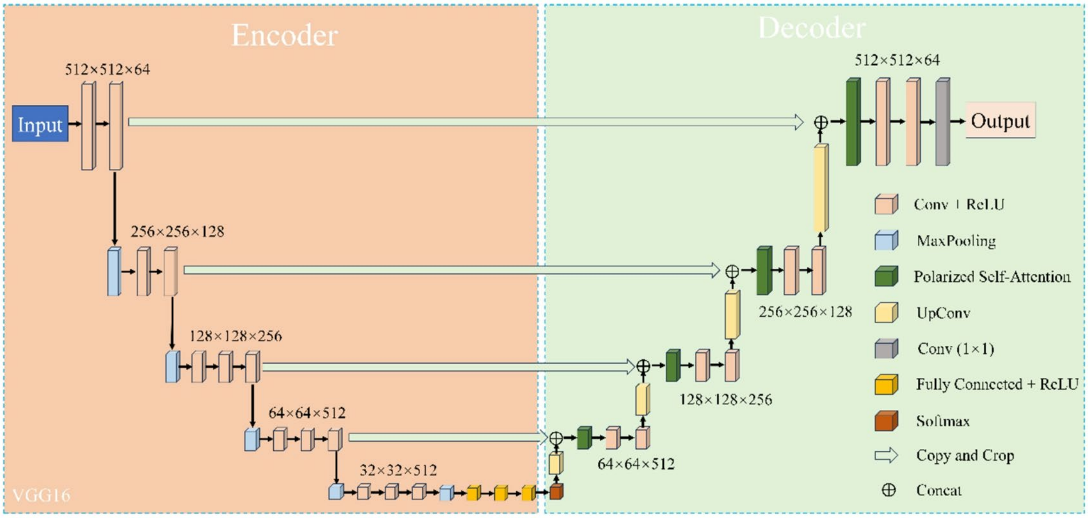

Fig. 3. Architecture of the VDP_UNet model.

UNet network

UNet is a classic semantic segmentation network that adopts a U-shaped encoder–decoder architecture[36]. The encoder is composed of multiple layers of convolution and max pooling, progressively extracting spatial and texture features from the image. Through successive 3 × 3 convolutions followed by ReLU activation, UNet enhances deep feature representation while preserving local information. Each downsampling operation halves the spatial resolution and increases the number of feature channels to capture more abstract high-level semantic information.

The decoder gradually restores the image resolution using transposed convolutions, and employs skip connections to pass shallow features from corresponding encoder layers[37], enabling effective feature fusion. This fusion aids in recovering fine details and improves segmentation accuracy. The fused features are further refined through 3 × 3 convolutions and finally processed by a 1 × 1 convolution to produce the semantic segmentation output.

VGG16

This study employs the structurally stable and widely used convolutional neural network VGG16 as the backbone for feature extraction[38]. VGG16 consists of multiple consecutive 3 × 3 convolutional layers, max pooling layers, several fully connected layers, and a Softmax classification layer, offering strong hierarchical feature representation capabilities. By stacking small-sized convolutional kernels, the network effectively expands the receptive field and enhances feature abstraction while controlling computational complexity, thereby improving its ability to capture fine image details. This architecture enables precise extraction of critical information—such as texture, edges, and shape—of plateau zokor mounds from high-resolution remote sensing images of alpine meadows. As a result, it significantly improves the model's ability to detect and segment small targets under complex background conditions.

Polarized self-attention module

Polarized Self-Attention (PSA) is a lightweight attention mechanism specifically designed for pixel-level regression tasks[39]. It splits the attention process into Channel-only Self-Attention and Spatial-only Self-Attention, modeling high-resolution attention separately along the channel and spatial dimensions. This design enables more precise capture of key structures and semantic information in images, significantly enhancing the expressiveness and discriminative power of feature representations.

In this study, PSA is implemented using a parallel structure that computes channel-only and spatial-only self-attention simultaneously. In the channel branch, the input feature map X is first passed through two 1×1 convolution layers to produce features q and v. The feature q is compressed to a single channel to extract a compact global representation, while v retains richer information with $C/2$ channels. The compressed q is then normalized using the Softmax function to highlight relative importance across channels. This normalized q is multiplied with the reshaped v to generate a channel-wise aggregated representation, which is then processed by another 1×1 convolution and LayerNorm to restore the original C dimensions. Finally, the attention weights are activated with a Sigmoid function, scaled to the [0,1] range, and multiplied channel-wise with the input feature map to enhance the output. In the spatial branch, two 1×1 convolutions are applied to extract a spatial feature map v and a globally averaged 1×1 feature q, which are used to compute spatial correlations. The resulting attention map is reshaped and passed through a Sigmoid activation, then multiplied pixel-wise with the input feature map to emphasize critical spatial regions. The overall architecture of the PSA module is shown in Fig. 4.

The computations for the Channel-only branch, Spatial-only branch, and the parallel arrangement of both branches are as follows:

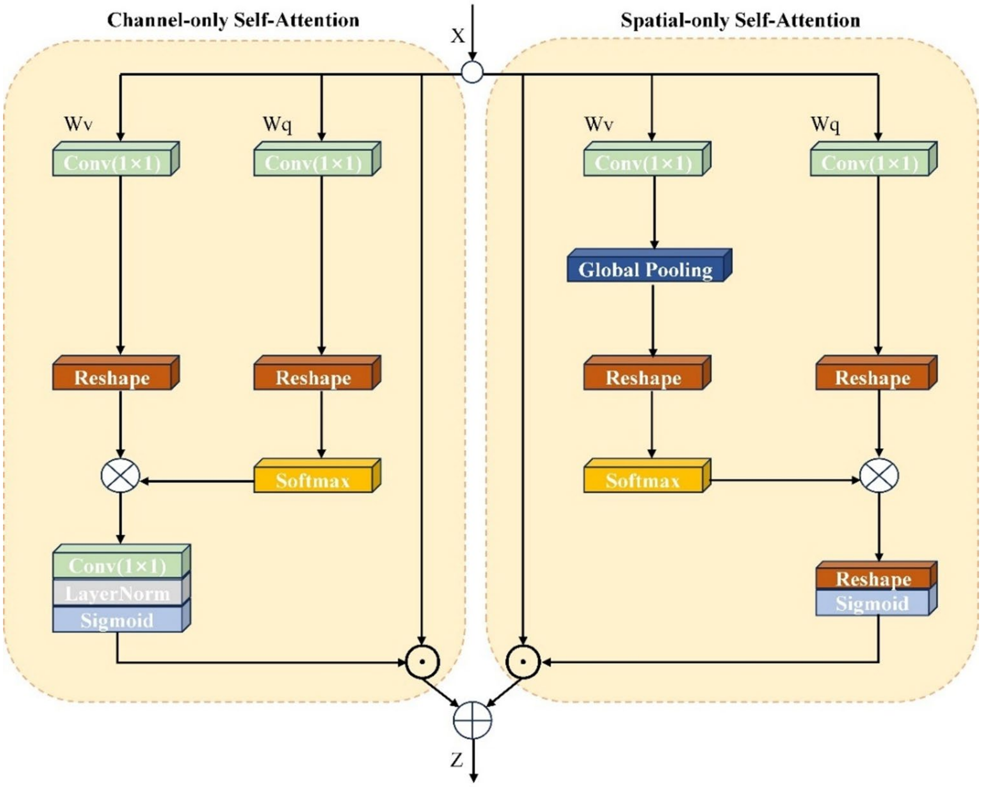

Fig. 4. Parallel architecture of polarized self-attention.

$$A^{ch}(X) = F_{SG}[W_{z\theta_1}((\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X))))] \tag{1}$$

$$A^{sp}(X) = F_{SG}[\sigma_3(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X)))] \tag{2}$$

$$PSA_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X)\odot^{ch}X + A^{sp}(X)\odot^{sp}X \tag{3}$$

where X is an input tensor, W_q, W_v and W_z are 1×1 convolution layers respectively, σ_1, σ_2 and σ_3 are tensor reshape operators, $F_{SG}()$ represents the attention-weighted operation on the value feature matrix, $F_{SM}(\cdot)$ is a softmax operator, "×" is the matrix dotproduct operation, $F_{GP}(\cdot)$ is a global pooling operator, where $\odot^{ch}$ is a channel-wise multiplication operator, where $\odot^{sp}$ is a spatial-wise multiplication operator and "+" is the element-wise addition operator.

Dice loss function

Dice Loss offers significant advantages in segmentation tasks involving small foreground objects and imbalanced class distributions[40]. Unlike cross-entropy loss, which calculates errors independently at each pixel and often overlooks small targets, Dice Loss directly optimizes the overlap between the predicted region and the ground truth, placing greater emphasis on overall structural consistency. This makes it particularly effective for accurately identifying small yet important plateau zokor mounds in alpine meadows. Additionally, Dice Loss helps mitigate the negative effects of class imbalance and enhances the model's ability to segment edge regions, significantly reducing contour blurring. Its formula is shown in Eq. (4):

$$Dice\ Loss = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \gamma}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \gamma} \tag{4}$$

Where p_i represents the predicted probability of the i-th pixel, y_i represents the ground truth label of the i-th pixel, N denotes the N-th pixel, and γ is a smoothing term used to prevent division by zero.

Parameter setting details

This study was conducted on a Windows 11 operating system using the PyTorch 2.0.0 deep learning framework. The server is equipped with an Intel(R) Core(TM) i9-14900 K processor and an NVIDIA GeForce RTX 4090 GPU, utilizing the CUDA v11.8 parallel computing platform and the cuDNN 8.9.7 deep neural network acceleration library. Python 3.8.20 was used as the programming language. Optimization was performed using the Adam optimizer with a momentum of 0.9 and a batch size of 16. Images were processed at a resolution of

512 × 512 pixels. The initial learning rate was set to 0.0001, with a minimum learning rate of 0.000001. The learning rate decay followed a cosine schedule, and training lasted for 80 epochs.

Evaluation metric

The model's performance was evaluated using key metrics derived from the confusion matrix: True Positives (TP) represent the number of pixels correctly identified as zokor mounds; False Positives (FP) are non-mound pixels mistakenly predicted as mounds; True Negatives (TN) refer to pixels correctly identified as non-mounds; and False Negatives (FN) are actual mound pixels incorrectly classified as non-mounds. Choosing appropriate evaluation metrics is critical for comprehensively assessing the effectiveness of the proposed mound extraction model. In this study, several essential metrics were used to evaluate semantic segmentation performance, including Intersection over Union (IoU), Mean Intersection over Union (MIoU), Mean Pixel Accuracy (MPA), Precision, Recall, Accuracy, F1-score, and FPS.

IoU measures the ratio of the overlap between the predicted and ground truth regions of a specific class to their union:

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

MIoU is used to measure the overlap between the predicted and actual zokor mound areas:

$$MIoU = \frac{1}{k+1} \cdot \sum\nolimits_{i=0}^{k} \frac{P_{ii}}{\sum\nolimits_{j=0}^{k} P_{ij} + \sum\nolimits_{j=0}^{k} P_{ji} - P_{ii}} \tag{6}$$

In Eq. (6), (k + 1) represents the number of categories. P_{ii} is the count of True Positive, P_{ij} represents False Negative, and P_{ji} indicates False Positive. In this context, '*i*' signifies the true category, while '*j*' refers to the other categories.

MPA is calculated by averaging the pixel accuracy for each class, where pixel accuracy refers to the ratio of correctly classified pixels of a given class to the total number of pixels in that class:

$$MPA = \frac{1}{C} \sum_{i=1}^{C} \frac{n_{ii}}{\sum\limits_{j} n_{ij}} \tag{7}$$

C represents the number of classes (for a binary classification task, *C* = 2), n_{ii} denotes the number of pixels correctly classified as class *i*, and n_{ij} is the total number of pixels belonging to class *i*.

Precision refers to the proportion of correctly classified zokor mound pixels among all pixels that were predicted to be zokor mounds:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall evaluates the ratio of correctly classified zokor mound pixels to the total number of pixels labeled as zokor mounds:

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Accuracy is used to measure the proportion of correctly classified pixels by the model at the pixel level:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

The F1-score, a key metric, is the harmonic mean of precision and recall; a higher F1-score indicates better model performance:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

FPS (Frames Per Second): Under the same hardware conditions, a higher FPS indicates stronger real-time processing capability of the model. FPS is calculated as: FPS = 1/latency, where latency refers to the time required for the network to process a single image.

Results

Ablation study

To validate the effectiveness of the proposed VDP_UNet method in segmenting zokor activity areas (zokor mounds) in alpine meadow regions, detailed ablation experiments were conducted on the constructed dataset. Various experimental configurations were designed to assess the impact of each improved module on the overall model performance and to quantify their contributions. Model performance was evaluated using IoU, MIoU, MPA, F1-score, and FPS. The corresponding experimental results are summarized in Table 1.

Case	VGG16	Dice Loss	PSA	IoU(%)	MIoU(%)	MPA(%)	F1-score(%)	FPS
1	-	-	-	47.12	65.62	73.19	50.29	33.56
2	√	-	-	48.38	73.82	78.43	65.21	30.42
3	√	√	-	50.35	74.79	81.84	66.98	31.07
4	√	-	√	48.84	74.06	78.01	65.63	28.78
5	√	√	√	**51.99**	**75.63**	**82.66**	**68.41**	**42.13**

Table 1. Ablation experiment results. Bold values indicate the best performance. A "√" denotes that the module is included, while a "-" indicates it is not.

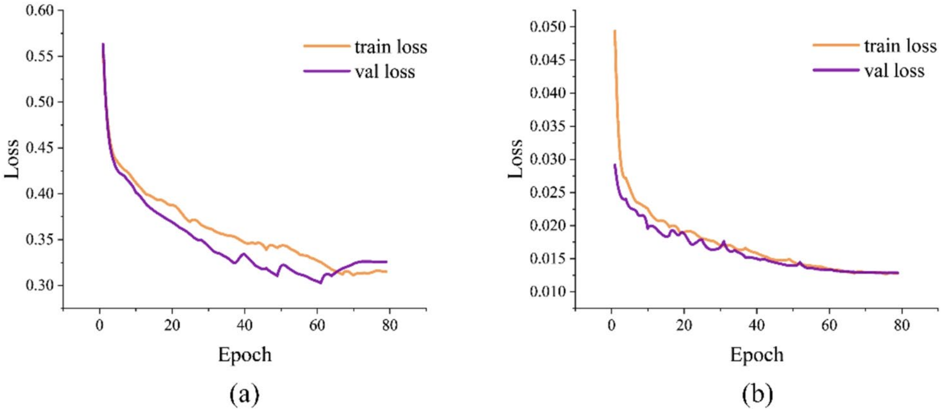

Fig. 5. Comparison of loss curves between the UNet (**a**) and VDP_UNet (**b**) models.

The experimental results demonstrate that in Case 2, integrating VGG16 alone led to improvements of 1.26% in IoU, 8.2% in MIoU, 5.24% in MPA, and 14.92% in F1-score compared to the baseline model, highlighting VGG16's advantage in extracting zokor mound features. The deeper architecture of VGG16 allows it to capture global contextual information from images, enabling the model to better focus on relevant mound features and improving its ability to accurately identify mound regions.

In Case 3, the Dice Loss function was added on top of the VGG16 backbone. By directly measuring the overlap between predicted results and ground truth labels, Dice Loss helps the model converge more effectively. In contrast, the baseline model uses cross-entropy with a predefined weight map, which struggles to emphasize hard examples and often overlooks contextual structural information. Compared to Case 2, Case 3 showed further increases of 1.97% in IoU, 0.97% in MIoU, 3.41% in MPA, 1.77% in F1-score, and 0.65 f/s in FPS.

Case 4 introduced the PSA module. The parallel structure of Channel-only and Spatial-only Self-Attention in PSA effectively enhances fine-grained features of zokor mounds in complex backgrounds and improves localization accuracy. Compared to Case 2, IoU, MIoU and F1-score increased by 0.46%, 0.24% and 0.42%, respectively, though MPA slightly decreased by 0.42%.

In Case 5, the combination of Dice Loss and the PSA module on the VGG16 backbone achieved the best overall performance, with IoU, MIoU, MPA, F1-score, and FPS reaching 51.99%, 75.63%, 82.66%, 68.41%, and 42.13f/s, respectively—improvements of 4.87%, 10.01%, 9.47%, 18.12%,and 8.57 f/s, over the baseline. These results validate the effectiveness of the proposed VDP_UNet model for zokor mound segmentation in the complex environment of alpine meadows.

The training and validation loss curves of the UNet and VDP_UNet models are shown in Fig. 5. As training progressed, the loss curves began to stabilize at epoch 70 and reached a steady state by epoch 80.

Comparative experiments

Comparison with classical algorithms

To ensure recognition accuracy, the proposed method was compared against several classic models on the alpine meadow zokor mound dataset, including SegFormer[41], DeepLabV3+[42], BiSeNetV2[43], Fast-SCNN[44], and DANet[45], All models were trained for 80 epochs.

As shown in Table 2, VDP_UNet outperforms several mainstream semantic segmentation models across key metrics such as Accuracy, MIoU, and F1-score. While SegFormer achieved the highest Precision (72.32%), its relatively low Recall and F1-score indicate an issue with under-detection, and therefore it was not selected as the baseline model. DeepLabV3 + achieved a slightly higher Recall than VDP_UNet, but fell short in other performance indicators. BiSeNetV2 and DANet demonstrated relatively balanced results across metrics and maintained stable overall performance, yet still underperformed compared to the proposed method. Fast-SCNN

Model	Precision(%)	Recall(%)	Accuracy(%)	MIoU(%)	F1-score(%)
SegFormer	**72.32**	41.18	99.10	67.33	52.48
DeepLabV3+	68.11	**66.12**	98.99	63.57	67.10
BiSeNetV2	60.84	58.39	98.86	57.06	59.59
Fast-SCNN	49.90	58.55	98.78	56.76	53.88
DANet	70.60	62.68	98.96	61.00	66.40
VDP_UNet	71.44	65.63	**99.27**	**75.63**	**68.41**

Table 2. Comparative analysis with classic semantic segmentation models. Bold font indicates the best values.

performed the worst on this dataset, struggling to effectively capture the small-scale features of zokor mounds. Overall, VDP_UNet strikes a strong balance between accuracy and robustness, confirming its effectiveness in extracting zokor mound regions under complex background conditions and offering valuable support and methodological insight for future research.

Figure 6 illustrates the loss curves of different semantic segmentation models. Although the SgFormer model achieved the highest accuracy, its loss decreased too rapidly and became stable around the 30th epoch, indicating that the model might suffer from an improper learning rate setting or premature convergence. The Fast-SCNN model showed relatively poor training performance, with large fluctuations in its loss curve throughout the training process and no clear convergence. Overall, the VDP_UNet model exhibited faster convergence and the lowest final loss value, demonstrating superior convergence and stability.

Backbone network comparison

This experiment compared the performance of models using different backbone networks—UNet's original backbone, VGG16, and ResNet50—on the zokor mound dataset. The experimental results are shown in Table 3. The results indicate that the VGG16 backbone demonstrates a stronger ability to capture contextual information from zokor mound images. It achieved Precision, IoU, and MIoU values of 76.04%, 48.38%, and 73.82%, respectively, representing improvements of 25.01%, 1.26%, and 8.20% over the original UNet backbone, and achieving the best overall performance.

Polarized self-attention

To verify the effectiveness of the Polarized Self-Attention (PSA) module in semantic feature fusion, a comparative experiment was conducted based on the VGG16 backbone using CBAM, Triplet, ECA, and PSA modules, as shown in Table 4. The PSA module integrates features through a parallel combination of channel and spatial attention mechanisms, achieving Precision, IoU, and MIoU values of 78.85%, 48.84%, and 74.06%, respectively—improvements of 2.81%, 0.46%, and 0.24% over the other modules. These results clearly demonstrate the superiority of the PSA module compared to the baseline modules, indicating that incorporating Polarized Self-Attention enables more effective integration of semantic information and enhances segmentation accuracy.

Loss function

A series of comparative experiments were conducted to verify the effectiveness of the proposed loss function. The experiments were carried out based on the VGG16 backbone and PSA module, as shown in Table 5. Three loss functions—Dice Loss, Focal Loss, and Dice + Focal—were introduced for comparison, where "Origin" represents the baseline model without additional loss function. The results show that the baseline model achieved IoU, MIoU, F1-score, and FPS values of 48.84%、74.06%、65.63%、28.78f/s, respectively. After introducing Dice Loss, these metrics improved to 51.99%, 75.63%, 68.41%, and 42.13 f/s, representing the best overall performance. These findings confirm the effectiveness of the proposed loss function in the zokor mound segmentation task and demonstrate its ability to significantly enhance feature extraction accuracy.

Result visualization

To visually demonstrate the performance of the VDP_UNet model in detecting zokor mound regions, two UAV images captured at an altitude of 30 m were randomly selected and enlarged for qualitative analysis. The comparison group included UNet, other comparative models, and manual annotations, resulting in a total of eight sets of experimental results, as shown in Fig. 7. In the figure, "Original" denotes the raw image, "Label" represents the manually annotated ground truth, and the black-and-white images indicate the model predictions, where white areas correspond to detected zokor mounds and black areas represent non-mound regions. The first sample was collected during the flourishing grass stage, while the second corresponds to the regreening stage, demonstrating the model's adaptability and detection performance across different vegetation growth phases.

As shown in Fig. 7, zokor mounds exhibit more distinctive visual features during the flourishing grass stage, especially in terms of color, making them easier to identify. In contrast, during the regreening stage, the mounds often overlap with ground objects such as livestock footprints and surface patches, resulting in less distinguishable features and a higher likelihood of confusion. Compared with the manually annotated results, DeepLabV3+, BiSeNetV2, Fast-SCNN, and DANet performed unsatisfactorily on the zokor mound test set—showing significant false detections and omissions during the flourishing grass stage and failing to accurately delineate mound boundaries in the complex backgrounds of the regreening stage. Although SegFormer achieved a slight improvement in detection accuracy, its boundary perception capability remained limited.

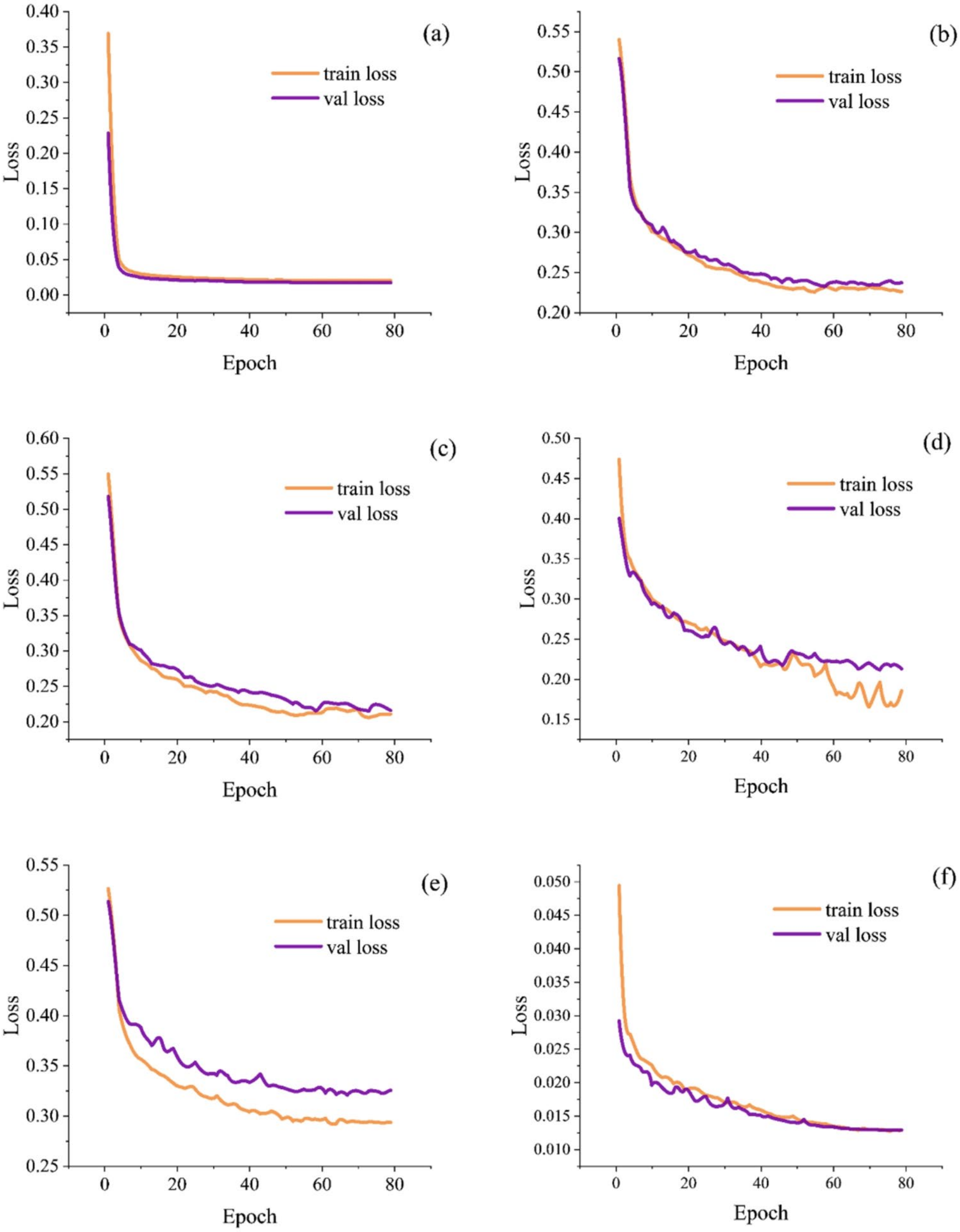

Fig. 6. Comparison of loss curves for different semantic segmentation models. (**a**) SegFormer model, (**b**) DeepLabV3 + model, (**c**) BiSeNetV2 model, (**d**) Fast-SCNN model, (**e**) DANet model, and (**f**) VDP_UNet model.

Method	Precision(%)	IoU(%)	MIoU(%)	Accuracy(%)	F1-score(%)	FPS(f/s)
Origin	51.03	47.12	65.62	99.17	50.29	**33.56**
ResNet50	67.57	48.08	73.61	99.18	64.90	30.81
VGG16	**76.04**	**48.38**	**73.82**	**99.26**	**65.21**	30.42

Table 3. Results from choosing the backbone, bold font indicates the best values.

In contrast, the proposed VDP_UNet exhibits superior detection performance across all four representative scenarios. This improvement is largely attributed to the VGG16 backbone, which enhances the model's capacity to capture contextual and semantic features of zokor mounds. Furthermore, the integration of a parallel PSA module—combining channel-only and spatial-only self-attention—significantly boosts the extraction of fine-grained details and edge information. The use of Dice Loss to address class imbalance further enhances label

Method	Precision(%)	IoU(%)	MIoU(%)	Accuracy(%)	F1-score(%)	FPS(f/s)
Origin	76.04	48.38	73.82	99.26	65.21	30.42
CBAM	76.80	47.58	73.42	99.26	64.48	**39.34**
Triplet	76.67	48.31	73.79	99.27	65.15	20.05
ECA	76.10	45.10	72.10	99.22	62.16	38.47
PSA	**78.85**	**48.84**	**74.06**	**99.29**	**65.63**	28.78

Table 4. The network performance is compared by using different network blocks, bold font indicates the best values.

Method	Precision(%)	IoU(%)	MIoU(%)	Accuracy(%)	F1-score(%)	FPS(f/s)
Origin	**78.85**	48.84	74.06	**99.29**	65.63	28.78
Dice + Focal	68.37	50.81	75.01	99.22	67.38	29.28
Focal Loss	77.94	48.08	73.65	99.26	64.90	41.14
Dice Loss	71.44	**51.99**	**75.63**	99.27	**68.41**	**42.13**

Table 5. The network performance is compared by adding different loss function, bold font indicates the best values.

prediction accuracy. Collectively, these enhancements demonstrate the robustness and reliability of VDP_UNet in accurately identifying zokor mound features under complex background conditions.

Application of the VDP_UNet model in field sites and area Estimation

To validate the effectiveness of the proposed method and explore its potential applications in zokor mound monitoring and sustainable ecosystem management, four zokor mound UAV images—randomly selected and excluded from training—were analyzed. A comparative study was conducted using traditional area calculation methods in ENVI (https://envi.geoscene.cn/). This study primarily calculated the total zokor mound area by counting the number of target pixels labeled as 1 (zokor mound) in the segmentation result maps, with non-zokor mound areas labeled as 0, then converting these pixel counts to actual ground area based on flight altitude and camera sensor parameters. The related area estimation principle is illustrated in Fig. 8.

The formula for calculating the actual area of zokor mounds is as follows:

$$Area_{zokor\ mound} = C_{pixel} \times GSD^2 = C_{pixel} \times (\frac{H \times SW}{f \times IW})^2 \tag{12}$$

Where C_{pixel} is the number of pixels, *GSD* is the ground sampling distance, *H* is the flight altitude (meters), *SW* is the sensor width (meters), *f* is the focal length (meters), and *IW* represents the image width (pixels).

As shown in Fig. 9, images (a)–(c) were captured during the returning green period, while image (d) was taken during the peak grass period. The ENVI software was used to perform supervised classification based on the maximum likelihood method, with regions of interest (ROIs) manually defined. Post-processing steps included clustering analysis and principal and minor component analysis. During the peak grass period, when zokor mounds exhibit distinct texture and color differences from the surrounding vegetation, both ENVI and the proposed VDP_UNet produce satisfactory segmentation results. However, in the returning green period, zokor mounds often blend with livestock manure, bare soil, and other ground features, resulting in a complex background. This complexity limits ENVI's ability to capture discriminative statistical features, leading to numerous false positives. In contrast, VDP_UNet maintains strong performance in this challenging setting, accurately identifying zokor mound regions with segmentation results that closely align with the original images, despite occasional misclassifications of non-mound areas.

Overall, ENVI's reliance on traditional statistical classification methods results in a complex processing workflow and issues like blurred boundaries and misclassification, revealing clear limitations in zokor mound area extraction. By comparison, VDP_UNet enables large-scale, efficient UAV image processing and demonstrates superior overall performance in comparative experiments. Its segmentation results closely align with actual zokor mound distribution, providing more accurate area estimates, thus validating VDP_UNet's applicability and effectiveness for zokor damage area extraction in alpine meadow environments.

Subsequently, 100 zokor mound images that were not used for training were analyzed. For each image, the ground truth mound area and the predicted value were calculated to generate a scatter plot, as shown in Fig. 10. The closer the points are to the reference line y = x, the more consistent the predicted values are with the ground truth.

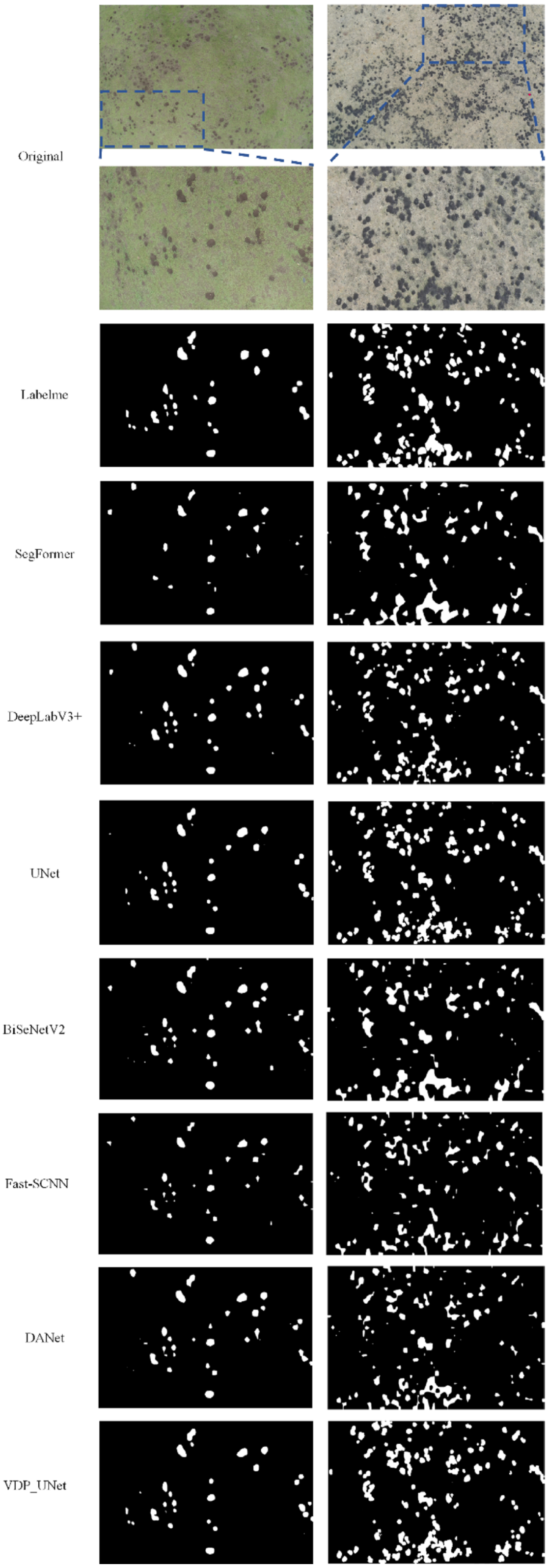

Fig. 7. Visualization results.

Discussion

Performance comparison of the VDP_UNet model with similar methods on the Zokor mound dataset

Semantic segmentation models have been widely applied across numerous object detection tasks, spanning

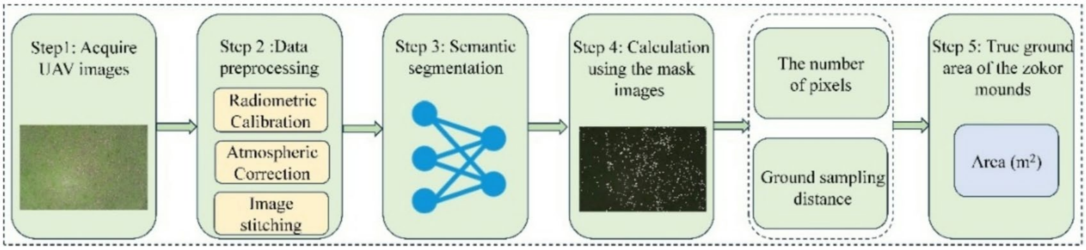
Step1: Acquire UAV images
Step 2 :Data preprocessing
Radiometric Calibration
Atmospheric Correction
Image stitching
Step 3: Semantic segmentation
Step 4: Calculation using the mask images
The number of pixels
Ground sampling distance
Step 5: True ground area of the zokor mounds
Area (m²)

Fig. 8. Workflow diagram for zokor mound area calculation in real field sites.

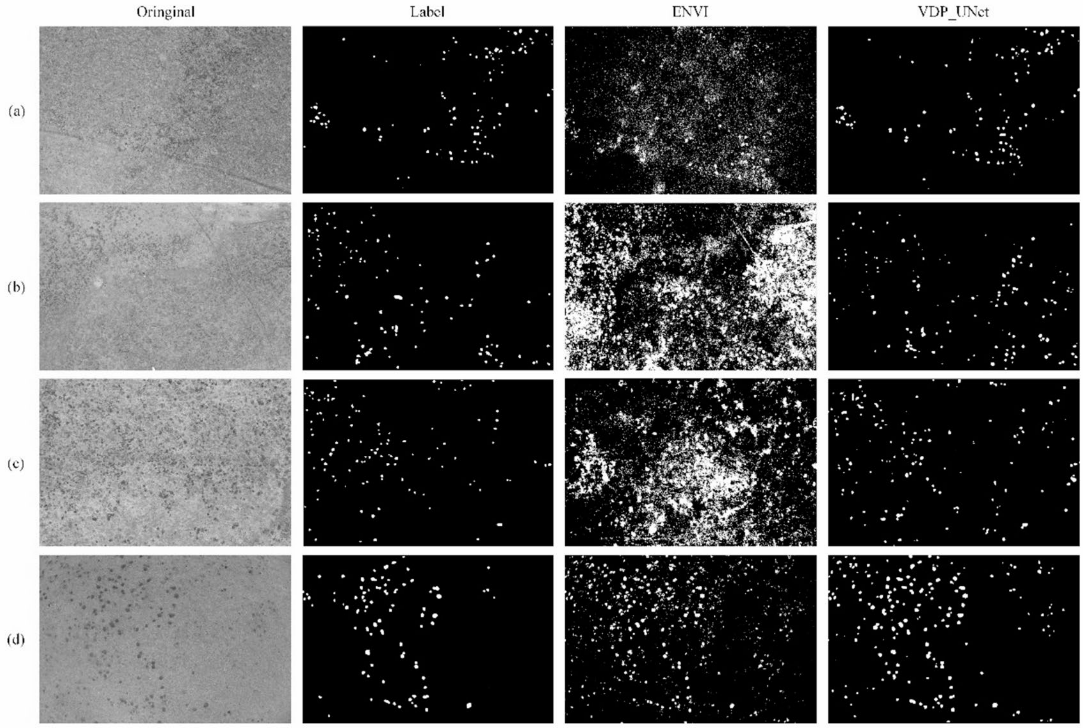

Fig. 9. Comparison of UAV images and segmentation results of zokor mounds in the real field using ENVI and VDP_UNet. "Label" refers to the precise manual annotations. Notes: ENVI's classification performance is poor with large errors; therefore, only images (**a**)–(**d**) segmented by VDP_UNet are shown, with the corresponding zokor mound areas in the field being 17.14 m², 20.47 m², 22.90 m², and 41.76 m², respectively.

fields such as agriculture, water resource management, forest fire early warning[46], and forestry monitoring[47], For example, Zhang et al.[48] used the DD-DA model for precise segmentation of gully erosion in the Northeast Black Soil region; Chen et al.[49] proposed Res_AUNet, which effectively enhanced the extraction of sunlight reflection areas on water surfaces; to address the efficiency bottlenecks and lack of explicit structural information fusion in traditional ViT models when processing high-resolution images, Chu et al.[50] designed a structurally improved Twins network architecture. Additionally, this study compared several mainstream semantic segmentation models, including M-DeepLabV3 + with a MobileNet backbone, V2-HRNet based on HRNetV2, and the lightweight M-PSPNet, a MobileNet-based variant of PSPNet. To comprehensively evaluate the performance of the proposed VDP_UNet model in zokor mound area extraction, all models were trained and tuned under identical conditions, including the same data splits, augmentation strategies, number of epochs, early stopping policy, learning rate schedules, and a consistent hyperparameter tuning budget, to ensure fairness and reliability of the comparison results.

At a flight altitude of 30 m, using UAVs to capture zokor mound images significantly improves data acquisition efficiency. However, it also presents challenges such as smaller target sizes and less distinguishable features, making mound extraction more difficult. To validate the effectiveness of the proposed method, we conducted comparative experiments between VDP_UNet and several mainstream semantic segmentation models tailored for small-object detection in UAV imagery, as shown in Table 6. The results demonstrate that VDP_UNet outperforms other models in terms of Precision (71.44%), MIoU (75.63%), Accuracy (99.27%), and F1-score (68.41%), achieving the best overall performance. The DD-DA model achieved the highest Recall (66.75%),

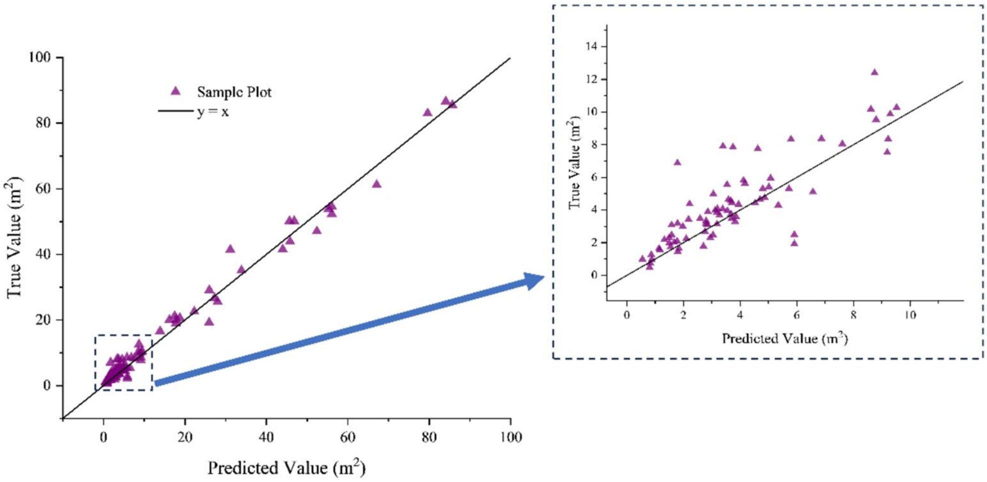

Fig. 10. Comparison between the ground truth and predicted values of zokor mounds in sample plots.

Model	Precision(%)	Recall(%)	MIoU(%)	Accuracy(%)	F1-score(%)
DD-DA	60.95	**66.75**	63.35	98.93	63.72
Res_AUNet	51.00	54.71	53.74	98.94	52.79
Twins	67.38	41.89	66.93	99.05	51.66
M-DeepLabV3+	59.07	58.54	70.32	99.01	58.80
V2-HRNet	63.21	56.60	70.82	99.08	59.72
M-PSPNet	46.21	41.18	63.28	98.71	43.55
VDP_UNet	**71.44**	65.63	**75.63**	**99.27**	**68.41**

Table 6. Performance comparison with existing semantic segmentation models. The bold font indicates the best values.

indicating strong capability in detecting zokor mounds. However, compared to DD-DA, VDP_UNet maintains high precision while also achieving competitive recall, highlighting its superior ability to capture fine-grained mound features.

Among the other models, Res_AUNet and M-PSPNet performed relatively poorly in terms of both precision and intersection over union, struggling to extract discriminative features of zokor mounds. Twins and V2-HRNet showed strengths in precision but suffered from lower recall, suggesting a higher risk of missed detections. Notably, introducing the PSA module between the encoder and decoder in Res_AUNet enhanced the model's representational power in complex backgrounds and improved its sensitivity to small-object semantic features. Nevertheless, there is still room for improvement in both its precision and recall metrics.

Limitations and future considerations

Although the proposed VDP_UNet model exhibited strong overall performance in this study, several limitations remain. Some missed detections were observed, and the model's precision still requires improvement. One contributing factor may be the limited availability of original zokor mound data, which could have hindered effective feature extraction and, consequently, model training. Additionally, the dataset used was confined to alpine meadows, which restricts the model's applicability to a relatively narrow ecological context. This study also focused exclusively on newly formed mounds, whereas a more comprehensive evaluation of rodent damage in grasslands should include semi-new and old mounds as well. Moreover, no comparison was made between the actual measured areas of zokor mounds and the area estimates generated by the proposed method, leaving a gap in assessing the model's practical utility.

Future research will focus on the following three directions. First, beyond visible light imagery, multimodal approaches that incorporate thermal infrared data, textual descriptions, and other complementary sources should be explored to enhance the extraction of zokor mound features across diverse modalities. Integrating various data types may improve the model's generalization capability in different environmental contexts. Second, while this study primarily demonstrated the feasibility of applying the segmentation model in alpine meadows, practical deployment remains limited. Future research could explore model lightweighting with respect to parameter count and model size, and incorporate boundary-aware loss and class-balance calibration to facilitate deployment on edge devices and support new application scenarios. Third, the scope of zokor mound research should be broadened to include different successional stages of mound development and to incorporate field-based area measurements. This includes, but is not limited to, identifying new, semi-new, and old mounds across diverse habitats such as alpine meadows and alpine shrub-meadows in regions like Qinghai, Tibet, Gansu, and Sichuan. These efforts will improve the applicability of mound area extraction methods, ensure greater consistency with real-world field conditions, and enhance both the model's generalizability and the reliability of experimental outcomes.

Conclusion

The widespread presence of zokor mounds poses a significant threat to the sustainable development of alpine meadow ecosystems. Accurately identifying their distribution not only facilitates the assessment of rodent damage but also serves as an indirect indicator of zokor activity intensity. To enhance the precision and efficiency of zokor mound extraction, this study proposes a deep semantic segmentation model—VDP_UNet—integrating a polarized self-attention mechanism. In VDP_UNet, the encoder is replaced with VGG16 to better capture global contextual information of mound regions. Additionally, a Polarized Self-Attention (PSA) block is introduced in the feature fusion stage following encoder-decoder skip connections to strengthen the representation of fine-grained features in complex backgrounds. The Dice loss function is employed to address sample imbalance and further improve overall model performance. Compared with classical semantic segmentation networks and several state-of-the-art methods, VDP_UNet achieves superior results across multiple evaluation metrics, demonstrating clear advantages in zokor mound extraction tasks. This approach provides a practical and effective solution for accurately detecting zokor mounds, offering strong potential for applications in rodent damage monitoring and ecological management. To support the real-world application of deep learning in this domain, a dedicated alpine meadow zokor mound dataset was constructed using UAV imagery collected at a 30-meter flight altitude. This dataset fills a critical gap in high-quality remote sensing data for zokor mounds and lays a solid foundation for future research and model development.

Data availability

1. The data that support the findings of this study are available in "Scicense Data Bank" at: [https://github.com/Yangyang875/Plateau-Zokor-Mounds] 2. The source code employed in the current research can be accessed on the GitHub page: [https://github.com/Yangyang875/VDP_UNet].

Received: 5 September 2025; Accepted: 4 November 2025
Published online: 12 December 2025

References

1. Wei, F. et al. Catalogue of mammals in China. *Acta Theriol. Sin.* **41** (5), 487–501. https://doi.org/10.16829/j.slxb.150595 (2021).
2. Tang, L. et al. Gene flows of Eospalax baileyi geographical populations Chinese. *J. Anhui Agricultural Sci.* **38** (10), 5123–5124. https://doi.org/10.13989/j.cnki.0517-6611.2010.10.033 (2010).
3. Feng, F., Gong, B. & Niu, K. Linking density of plateau Pika to vegetation characteristics and soil attributes in response to different grazing RegimesChinese. *Pratac Sci.* **36** (11), 2915–2925. https://doi.org/10.11829/j.issn.1001-0629.2019-0136 (2019).
4. Zhang, Z. et al. Grazing reduced vegetation biomass and root nutrition related to plateau Zokor creating mounds in summer on the Tibetan plateau. *Ecol. Eng.* **209**, 107404. https://doi.org/10.1016/j.ecoleng.2024.107404 (2024).
5. Su, J. et al. Zokor disturbances indicated positive soil microbial responses with carbon cycle and mineral encrustation in alpine grassland. *Ecol. Eng.* **144** (2), 105702–105702. https://doi.org/10.1016/j.ecoleng.2019.105702 (2020).
6. Yue, D. et al. Soil wind erosion and nutrient loss in typical rodent mounds in a degraded alpine grassland in the yellow river source ZoneChinese. *Arid Zone Res.* **41** (04), 603–617. https://doi.org/10.13866/j.azr.2024.04.07 (2024).
7. Zhang, X. & Li, G. Effects of rodents activities on grazing land and ecosystem in alpine MeadowChinese. *Pratac Sci.* **32** (05), 816–822. https://doi.org/10.11829/j.issn.1001-0629.2014-0016 (2015).
8. Hua, L. & Chai, S. Rodent pest control on grasslands in china: current state, problems and prospects. *J. Plant. Prot.* **49** (01), 415–423. https://doi.org/10.13802/j.cnki.zwbhxb.2022.2022812 (2022).
9. Pu, Q., Wang, Z., Hou, Q., Zhang, Z. & Su, J. Interspecific difference of mound morphological characteristics of two Zokor species in the Eastern margin of the Qilianshan mountains. *Pratac Sci.* **41** (05), 1221–1231. https://doi.org/10.11733/j.issn.1007-0435.2023.06.029 (2024).
10. Wen, X. et al. Research progresses in occurrence status and control technology of forest rodents in China. *World Forestry Res.* **34** (02), 91–95. https://doi.org/10.13348/j.cnki.sjlyyj.2020.0104.y (2021).
11. Qin, H. et al. Soil degradation intensity and Spatial distribution characteristics of alpine grassland in Qinghai Province. *Acta Agrestia Sin.* **29** (S1), 104–112. https://doi.org/10.11733/j.issn.1007-0435.2021.Z1.012 (2021).
12. Wang, Z. et al. Quantitative assess the driving forces on the grassland degradation in the Qinghai–Tibet Plateau, in China. *Ecol. Inf.* **33**, 32–44. https://doi.org/10.1016/j.ecoinf.2016.03.006 (2016).
13. Liu, C. *Grass Protection* (China Forestry Publishing House, 2009).
14. Plaza, J. et al. Classification of airborne multispectral imagery to quantify common vole impacts on an agricultural field. *Pest Manage. Sci.* **78** (6), 2316–2323. https://doi.org/10.1002/ps.6857 (2022).
15. Tuomi, M. W. et al. Novel frontier in wildlife monitoring: identification of small rodent species from fecal pellets using near-infrared reflectance spectroscopy (NIRS). *Ecol. Evol.* **13** (3). https://doi.org/10.1002/ece3.9857 (2023). e9857.
16. Du, B. et al. Analysis of Spatial distribution characteristics and influencing factors of lasiopodomys Brandtii based on unmanned aerial vehicle (UAV) low altitude remote sensing images. *Acta Ecol. Sin.* **45** (03), 1494–1502. https://doi.org/10.20103/j.stxb.202306071206 (2025).
17. Hua, L., Chu, B., Zhou, Y., Ma, S. & Zhou, J. .w. A method for investigating mounds number of subterranean rodent and its distribution. *Chin. J. Zool.* **53** (03), 461–467. https://doi.org/10.13859/j.cjz.201803014 (2018).
18. Fang, Y. & Shi, Q. Discussion on the hazards and prevention of rats and rabbits in Gansu grassland plateau. *Gansu Agric.* **06**, 41–42 (1998). https://doi.org/CNKI:SUN:GSNY 0.1998-06-013.
19. Wang, L. et al. Study on UAV monitoring and evaluation techniques for plateau Zokor damage. *Grassl. Turf.* **44**, 1–14 (2024).
20. Hua, R. et al. Monitoring of rodent damage areas in grassland using unmanned aerial vehicle remote sensing technology. *Acta Prataculturae Sin.* **32** (05), 71–82. https://doi.org/10.11686/cyxb2022234 (2023).
21. Karaucak, M., Steiniger, D. & Boroffka, N. A remote sensing-based survey of archaeological/heritage sites near Kandahar, Afghanistan through publicly available satellite imagery. *PLoS ONE.* **16** (11), e0259228. https://doi.org/10.1371/journal.pone.0259228 (2021).

22. Monsimet, J. et al. UAV data and deep learning: efficient tools to map ant mounds and their ecological impact. *Remote Sens. Ecol. Conserv.* **11** (1), 5–19. https://doi.org/10.1002/rse2.400 (2025).
23. Qi, G., Yu, Z., Shan, Y. & Tian, Y. Identification of rodent hole patches in desert grasslands using UAV imagery and OBIA-CFS algorithms. *Pratac Sci.* **41** (10), 2275–2283. https://doi.org/10.11829/j.issn.1001-0629.2023-0244 (2024).
24. Li, P. et al. Estimating area of grassland rodent damage rangeland and rat wastelands based on remote sensing in Altun Mountain, Xinjiang, China. *Xinjiang Agric. Sci.* **53** (07), 1346–1355. https://doi.org/10.6048/j.issn.1001-4330.2016.07.023 (2016).
25. Sandino, J., Wooler, A. & Gonzalez, F. Towards the automatic detection of pre-existing termite mounds through UAS and hyperspectral imagery. *Sens* **17** (10), 2196. https://doi.org/10.3390/s17102196 (2017).
26. Chen, J., Fan, J., Zhang, M., Zhou, Y. & Shen, C. MSF-Net: A multiscale supervised fusion network for Building change detection in high-resolution remote sensing images. *IEEE Access.* **10**, 30925–30938. https://doi.org/10.1109/ACCESS.2022.3160163 (2022).
27. Du, M., Wang, D., Liu, S. & Lv, C. Zhu, Y.p. Rodent hole detection in a typical steppe ecosystem using UAS and deep learning. *Front. Plant. Sci.* **13**, 992789. https://doi.org/10.3389/fpls.2022.992789 (2022).
28. Guo, X. et al. Study on key problems for rat hole recognition and count near ground based on deep learning and its application. *Acta Agric. Univ. Zhejiangensis.* **36** (09), 2146–2154. https://doi.org/10.3969/j.issn.1004-1524.20230858 (2024).
29. Sullivan, T. P. & Sullivan, D. S. Forecasting vole population outbreaks in forest plantations: the rise and fall of a major mammalian pest. *Ecol. Manage.* **260** (6), 983–993. https://doi.org/10.1016/j.foreco.2010.06.017 (2010).
30. Minaee, S. et al. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968 (2022).
31. Liu, J., Dong, R., Hua, L., Chu, B. & Ye, G. Effects of plateau Zokor Eospalax baileyi disturbance on the aboveground-belowground biomass allocation pattern of plant communities and soil physicochemical properties in alpine meadow. *J. Plant. Prot.* **51** (05), 1090–1098. https://doi.org/10.13802/j.cnki.zwbhxb.2024.2024815 (2024).
32. Zhang, J. et al. Plant functional group characteristics in different successional stages of plateau Zokor mounds and influencing factors. *J. Plant. Prot.* **51** (05), 1078–1089. https://doi.org/10.13802/j.cnki.zwbhxb.2024.2024814 (2024).
33. Bao, G. et al. Effects of plateau Zokor burrowing activity on soil nutrient Spatial heterogeneity in alpine grasslands. *Acta Prataculturae Sin.* **25** (07), 95–103. https://doi.org/10.11686/cyxb2015560 (2016).
34. Liu, S. et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision*. 38–55 (2024). https://doi.org/10.48550/arXiv.2303.05499
35. Zhang, H. et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *Comput. Vis. Pattern Recognit.* https://doi.org/10.48550/arXiv.2203.03605 (2022).
36. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. *Med. Image Comput. computer-assisted intervention–MICCAI 2015: 18th Int. Conf. Munich Ger. Oct. 5–9 2015 Proc. part. III*. **18**, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28 (2015).
37. Wu, D., Wang, Y., Xia, S. T., Bailey, J. & Ma, X. Skip connections matter: on the transferability of adversarial examples generated with Resnets. *Mach. Learn.* https://doi.org/10.48550/arXiv.2002.05990 (2020).
38. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *Comput. Vis. Pattern Recognit.* https://doi.org/10.48550/arXiv.1409.1556 (2014).
39. Liu, H., Liu, F., Fan, X. & Huang, D. Polarized self-attention: towards high-quality pixel-wise mapping. *Neurocomputing* **506**, 158–167. https://doi.org/10.1016/j.neucom.2022.07.054 (2022).
40. Li, X. et al. Dice loss for Data-imbalanced NLP tasks. 465–476 (2020). https://doi.org/10.18653/v1/2020.acl-main.45
41. Xie, E. et al. SegFormer: simple and efficient design for semantic segmentation with Transformers. *Adv. Neural. Inf. Process. Syst.* **34**, 12077–12090. https://doi.org/10.48550/arXiv.2105.15203 (2021).
42. Yang, Z., Peng, X., Yin, Z. & Yang, Z. Deeplab_v3_plus-net for image semantic segmentation with channel compression. *2020 IEEE 20th Int. Conf. Communication Technol. (ICCT)*. **1320-1324** https://doi.org/10.1109/ICCT50939.2020.9295748 (2020).
43. Yu, C. et al. BiSeNet V2: bilateral network with guided aggregation for real-time semantic segmentation. *Int. J. Comput. Vision*. **129** (11), 3051–3068. https://doi.org/10.1007/s11263-021-01515-2 (2021).
44. Poudel, R. P. K., Liwicki, S. & Cipolla, R. J. a.e.-p., 2019. fast-SCNN: fast semantic segmentation network. arXiv:1902.04502. https://doi.org/10.48550/arXiv.1902.04502
45. Fu, J. et al. Dual attention network for scene segmentation. *ArXiv:1809 02983*. https://doi.org/10.48550/arXiv.1809.02983 (2018).
46. Feng, H. et al. U3UNet: an accurate and reliable segmentation model for forest fire monitoring based on UAV vision. *Neural Netw.* **185**, 107207. https://doi.org/10.1016/j.neunet.2025.107207 (2025).
47. Zhang, J. et al. Detecting pest-infested forest damage through multispectral satellite imagery and improved UNet+. *Sensors* **22** (19). https://doi.org/10.3390/s22197440 (2022).
48. Zhang, X. et al. Remote sensing image segmentation of gully erosion in a typical black soil area in Northeast China based on improved DeepLabV3 + model. *Ecol. Inf.* **84**, 102929. https://doi.org/10.1016/j.ecoinf.2024.102929 (2024).
49. Chen, J. et al. Detecting sun Glint in UAV RGB images at different times using a deep learning algorithm. *Ecol. Inf.* **81**, 102660. https://doi.org/10.1016/j.ecoinf.2024.102660 (2024).
50. Chu, X. et al. Twins: revisiting spatial attention design in vision transformers. (2021).

Acknowledgements

The authors thank the reviewers for the helpful comments.

Author contributions

Y.Y.: Writing–original draft, data curation, funding acquisition, and conceptualization. L.W.: Writing–review and editing, conceptualization, methodology, project administration and supervision and. L.H.: Writing–review and editing, project administration, supervision, and funding acquisition.

Funding

This work was supported by the Natural Science Foundation of Sichuan Province [grant number 2023NSFSC0207], the University Scientific Research and Innovation Team Program of Sichuan [grant number 2024CXPT-07], and the Graduate Research and Innovation Projects of Gansu Province [grant number 2025CXZX-798].

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.W. or L.H.

Reprints and permissions information is available at www.nature.com/reprints.