Improving Geological Remote Sensing Interpretation Via a Contextually Enhanced Multiscale Feature Fusion Network

Kang He, Zhijun Zhang, Yusen Dong, Depan Cai, Yue Lu, and Wei Han

***Abstract*—Geological remote sensing interpretation plays a pivotal role in the field of regional geological mapping, encompassing the analysis of rock, soil, and water features. However, these geological elements can be obscured by the surrounding geographical environment and can undergo modifications caused by geological activities. The former hinders the effectiveness of satellite remote sensing data, resulting in the invisibility of element features, while the latter leads to the complex distribution of element features and significant spatial variations of geological elements. Consequently, existing deep learning-based models for interpreting geological elements often exhibit limited accuracy. To address these issues, this study proposes the contextually enhanced multiscale feature fusion network for the efficient interpretation of geological elements. First, the context enhancement module is employed to extract abundant feature information and reinforce contextual features, aiming to capture essential features and strengthen their interconnections. Second, the multiscale feature fusion module incorporates the SimAM attention mechanism to adaptively learn features from different channels, emphasizing the feature information that contributes to interpretation results and maximizing the comprehensive and crucial feature information for each element. Extensive experiments demonstrate the superior performance of both the context enhancement module and the multiscale feature fusion module compared to several representative deep learning networks in terms of overall interpretation accuracy on two datasets. The model demonstrated improvements in oPA and mIoU of 2.4% and 2.8%, respectively, on the Landsat 8 dataset, and 3.5% and 3.2%, respectively, on the Sentinel-2 dataset.**

***Index Terms*—Deep learning, feature fusion, geological remote sensing, semantic segmentation.**

Manuscript received 7 September 2023; revised 27 December 2023; accepted 1 March 2024. Date of publication 12 March 2024; date of current version 23 May 2024. This work was supported in part by the Geological Survey projects conducted by the Geological Survey of China under Grant DD20220995, Grant DD20230135, and Grant ZD20220409 and in part by the National Natural Science Foundation of China under Grant U21A2013 and Grant 41925007. *(Corresponding author: Yusen Dong.)*

Kang He and Yue Lu are with the Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences Wuhan, Wuhan 430078, China (e-mail: kang_kang_he@163.com).

Zhijun Zhang is with the Key Laboratory of Geological Survey and Evaluation of Ministry of Education, China University of Geosciences Wuhan, Wuhan 430078, China, and with the Xining Center of Natural Resources Comprehensive Survey, China Geological Survey, Xining 810000, China.

Yusen Dong is with the Key Laboratory of Geological Survey and Evaluation of the Ministry of Education, School of Computer Science, Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, China (e-mail: dongyusen@gmail.com).

Depan Cai and Wei Han are with the School of Computer Science and Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430078, China.

Digital Object Identifier 10.1109/JSTARS.2024.3374818

I. Introduction

THE interpretation of geological elements related to rock, soil, and water is a vital aspect of remote sensing in the geological environment. It plays a significant role in diverse domains such as geological mapping [1], [2], evaluation of off-road trafficability [3], selection of outdoor construction sites [4], [5], and early warning systems for natural disasters [6]. Currently, the interpretation of geological elements such as rocks, soil, and water still relies primarily on traditional methods such as visual interpretation by experts and field investigations [7]. The method of manual visual interpretation involves visually analyzing the characteristics of terrain imagery, amalgamating expert knowledge gleaned from geology and geography, and integrating comprehensive analysis and logical reasoning with other nonremote sensing data, thus accomplishing the meticulous interpretation of geological elements like rocks, soil, and water with high precision.

Manual visual interpretation, requiring staff to possess a myriad of comprehensive professional knowledge, is labor-intensive, characterized by diminished efficiency, and marked by substantial subjectivity. As remote sensing image data continues to grow exponentially, manual visual interpretation struggles to meet the modern society's demands for remote sensing information applications. With the widespread application of machine learning techniques, scientists have proposed numerous effective methods that have found extensive use in geological remote sensing, yielding remarkable achievements. These achievements encompass tasks such as water resource extraction, land cover interpretation, and landslide identification [8], [9]. Inspired by these advancements, some studies have extended these methods to the task of geological element interpretation. These methods can be categorized into models based on machine learning (ML) and models based on deep learning (DL) [10], [11], [12].

In contrast to manual visual interpretation, most ML methods have yielded substantial achievements in the interpretation of geological elements, capitalizing on their robust feature extraction and representation capabilities [13]. However, owing to the models' limited capacity in handling intricate nonlinear equations, the classification accuracy of ML models often falls short [14]. DL-based models, evolving from neural networks, harness the potency of extensive training data and exploit the depth of numerous hidden layers to acquire more meaningful features [15]. This process entails training a resilient nonlinear

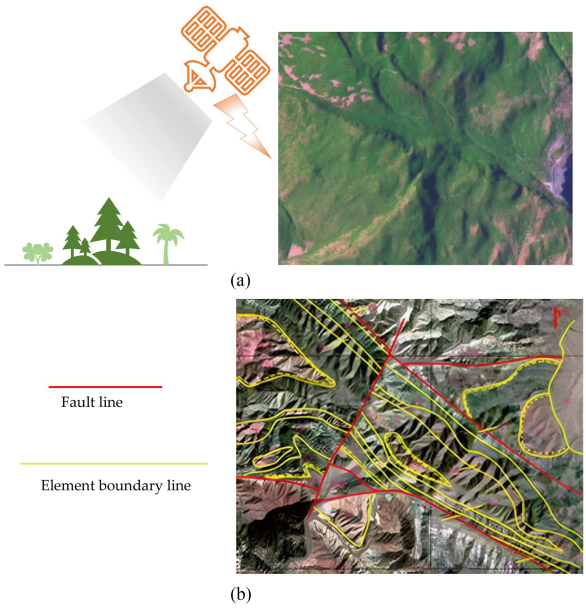

Fig. 1. Impediments of geographical environment and geological activities on the interpretation of geological elements. (a) Coverage scenarios of geographical environment. (b) Complex scenarios of geological activities.

deep neural network and subsequently deploying the trained model for predictive outputs. A multitude of research outcomes have validated the preeminent role of DL in the realm of image classification, at times even surpassing human capacity for task resolution [16]. For instance, DL-based models like CNNs [17], [18], [19], VGG [20], UNet [21], [22], and transformers [23] have achieved superior outcomes in interpreting geological elements like rock, soil, and water, surpassing the results obtained through conventional methods. Nonetheless, there remains substantial scope for enhancing these approaches in the pursuit of geological element interpretation.

In the realm of geological element interpretation using DL models, remarkable performance has been attained in completely exposed regions, while accuracy in other areas exhibits notably diminished precision. This phenomenon arises from the intricate geological scenarios in reality, where geological elements such as rock, soil, and water are susceptible to being concealed by vegetation and can undergo modifications due to geological activities [24]. The coverage of geographical environments adversely affects the observability of satellite remote sensing data, resulting in challenges such as object occlusion and the invisibility of geological element features. For instance, the surfaces of rocks, soil, and water can be covered by vegetation, leading to visual homogeneity across most areas of remote sensing images and a high degree of similarity between categories, as depicted in Fig. 1(a). Geological activities contribute to the complex spatial distributions of object features, characterized by fragmented structures and the significant spatial variability of geological elements [25], [26]. Faults and joints, for example, disrupt the initially continuous and concentrated distribution of rocks and soil, resulting in scattered fragments of various sizes, as depicted in Fig. 1(b). Folds also play a significant role in altering the spatial distribution of geological elements. Moreover, due to geological activities such as weathering, erosion, and deposition, the same rock type can exhibit distinct image features after fragmentation, showcasing intraclass dissimilarity [27], [28]. Prior research has tended to overlook the challenges inherent in genuine geological environment scenarios, failing to concentrate on extracting these concealed features and portraying intricate structural characteristics. Within this study, it has been observed that appropriately guiding the model toward uncovering the pivotal yet concealed attributes of geological elements like rock, soil, and water, while concurrently establishing interconnections among the elemental features, can significantly enhance the performance of geological element interpretation.

To tackle the aforementioned challenges, this study proposes a novel network model called the contextually enhanced multiscale feature fusion network (CEMFFNet), which integrates context enhancement extraction (CEE) and multiscale feature fusion (MFF). The CEE module is designed to extract abundant feature information and strengthen contextual features, enabling the capture of crucial features and enhancing the interconnections between them [29]. The MFF module incorporates the SimAM attention mechanism to facilitate adaptive feature learning across different channels, effectively emphasizing the feature information that significantly affects interpretation outcomes [30]. Consequently, this approach maximizes the acquisition of comprehensive and pivotal feature information regarding the target objects. The contributions of this study are primarily manifested in three key aspects:

1) The research area selected for this study encompasses a region in Iran characterized by typical geological conditions. Two satellite image datasets were created specifically for this study: the Landsat 8 dataset with a resolution of 15 m and the Sentinel-2 dataset with a resolution of 10 m.
2) In the context of geological element interpretation, an innovative network model, CEMFFNet, is introduced. This model is custom-designed to tackle the challenge of geological element interpretation and is employed for the interpretation of rock, soil, and water elements using Landsat 8 and Sentinel-2 images within the designated study area.
3) The proposed CEMFFNet model, through the synergistic integration of the context enhancement module and the MFF module, achieves proficient extraction and fusion of crucial features in geological element interpretation. This augmentation significantly enhances the accuracy of geological element interpretation for rock, soil, and water elements.

The rest of this article is organized as follows: Section II introduces the related work, Section III introduces the research area and dataset, Section IV outlines the methodology, Section V elaborates on the experimental results, and Finally, Section VI presents the conclusion.

II. Related Work

Geological element interpretation is a significant scientific pursuit directed toward the comprehensive classification of rock, soil, and water types within the study region. Manual visual

interpretation techniques entail substantial labor and resources, coupled with a deficiency in immediacy and objectivity. As a consequence, an expansive range of ML and DL methodologies has been harnessed for the automated interpretation of geological elements. A retrospective scrutiny of these two genres of approaches is presented herein.

The initial methods for automated geological interpretation of rock, soil, and water elements involved the utilization of ML techniques to construct mathematical models for accomplishing interpretation tasks. Bachri et al. [31] implemented a support vector machine (SVM) interpretation technique that integrated geomorphological features with Landsat 8 OLI multispectral data to map lithological units in the Souk Arbaa Sahel area of the Sidi Ifni River, located in the western Anti-Atlas region of southern Morocco. Bressan et al. [32] applied ML methods to classify lithology using multivariate log data from offshore oil wells in the International Ocean Discovery Program, demonstrating the potential of ML in enhancing geological research. Kumar et al. [33] introduced an automated lithological mapping method in the Hutti region of India's greenstone belt, employing airborne visible infrared imaging spectrometer-next generation hyperspectral data. Through the use of spectral enhancement techniques and ML algorithms, they successfully generated high-resolution reference lithological maps.

ML methods commonly initiate from predetermined features and expert regulations, subsequently utilizing ML models to train classifiers for accurately discerning diverse geological components like rock, soil, and water. However, these ML-based models confront certain nonlinear constraints in capturing intricate feature representations from the training samples, which could result in accuracy levels that fall short of expectations.

DL methodologies leverage their robust nonlinear adaptability and feature learning transformation capabilities, achieving remarkable feats across various remote sensing tasks in recent years [34]. These methodologies can automatically learn spectral and spatial information from remote sensing and geographical data, translating them into advanced feature representations [35], [36]. As a result, they dominate image processing fields like image classification, semantic segmentation, and object detection. DL methods are employed to compensate for the limitations of ML methods, and a plethora of research outcomes have substantiated its superiority in remote sensing classification tasks.

Methods based on DL predominantly approach the task of geological element interpretation from the perspective of semantic segmentation. Liu et al. [37] proposed a framework that combined multisource data fusion techniques with a fully convolutional network (FCN) model to enhance the accuracy of geological mapping. The results showcased the framework's efficacy in accurately identifying geological features, including lithological units. Shirmard et al. [38] employed convolutional neural networks (CNNs) and traditional ML methods, including SVM and multilayer perceptrons, to map lithological units in a mineral-rich region in southeastern Iran. The results demonstrated that both CNNs and traditional ML methods were effective in utilizing their respective remote sensing data sources to generate accurate lithological maps of the study area. Han et al. [39] proposed an adaptive multisource data fusion network (AMSDFNet), which leveraged DL features, to enhance the interpretation of geological elements, including rock, soil, surface water, and glaciers. The AMSDFNet was specifically designed to efficiently interpret multiple geological remote sensing elements by harnessing the power of DL methods. Wang et al. [40] applied CNNs to map lithological units using geochemical data from fluvial sediments in the Daqiao gold mining area of the Western Qinling Mountains, demonstrating the potential of CNN in geological mapping by accurately delineating 15 trace elements within the study area. To enhance the precision of intelligent soil element interpretation, Lu et al. [23] introduced an implicit-knowledge-guided adaptive feature fusion network (IAFFNet). Their study highlighted the latent capacity of implicit knowledge in soil element interpretation. However, in genuine geological settings, these methods have disregarded the impact of geographical coverage and geological activities on the initial remote sensing images as well as the spatial arrangement of geological elements such as rock, soil, and water.

III. Research Area and Dataset

A. Research Area

The research area, illustrated in Fig. 2, is situated in the northern coastal region of the Strait of Hormuz, within the territorial confines of Iran. Iran, a Middle Eastern nation, is located in Western Asia, bordered by the Caspian Sea to the north and the Persian Gulf and Arabian Sea to the south. The northern coastline of the Caspian Sea and the southwestern coastline of the Persian Gulf showcase small portions of alluvial plains. This area is influenced by a temperate continental climate and a highland mountain climate, renowned for its diverse array of soil types [41]. The predominant rock formations in the region consist of Quaternary volcanic rocks and Mesozoic greenstones. The surface topography exhibits a combination of coarse-grained soil, rock formations, and deposits of fine-grained soil and gravel. Gradually, soil salinity increases from rocky mountainous areas to beach surfaces. Both topography and climate play pivotal roles in the genesis of geological materials in this region [42]. Given the proximity of the study area to the coast and shoreline, it is imperative to discern the material composition and geological genesis types, as this knowledge is indispensable for national spatial planning, comprehensive environmental preservation, systematic restoration, and holistic management.

Nevertheless, the study area encompasses an expansive territory with a diverse array of rock formations and soil characteristics. The conventional methodology of "manual visual interpretation + field surveys" proves to be both laborious and inefficient for large-scale operations. Hence, harnessing the potential of remote sensing technology to swiftly capture real-time image data, coupled with the utilization of DL techniques for efficient and expeditious interpretation of rock, soil, and water features, holds immense significance in meeting the burgeoning requirements of remote sensing applications.

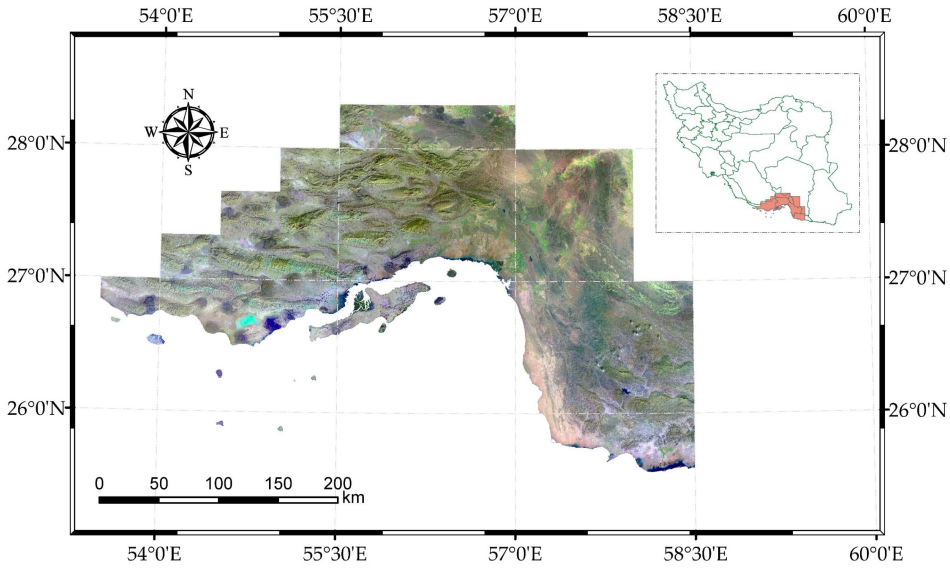

Fig. 2. Geographical location map of the research area in Iran.

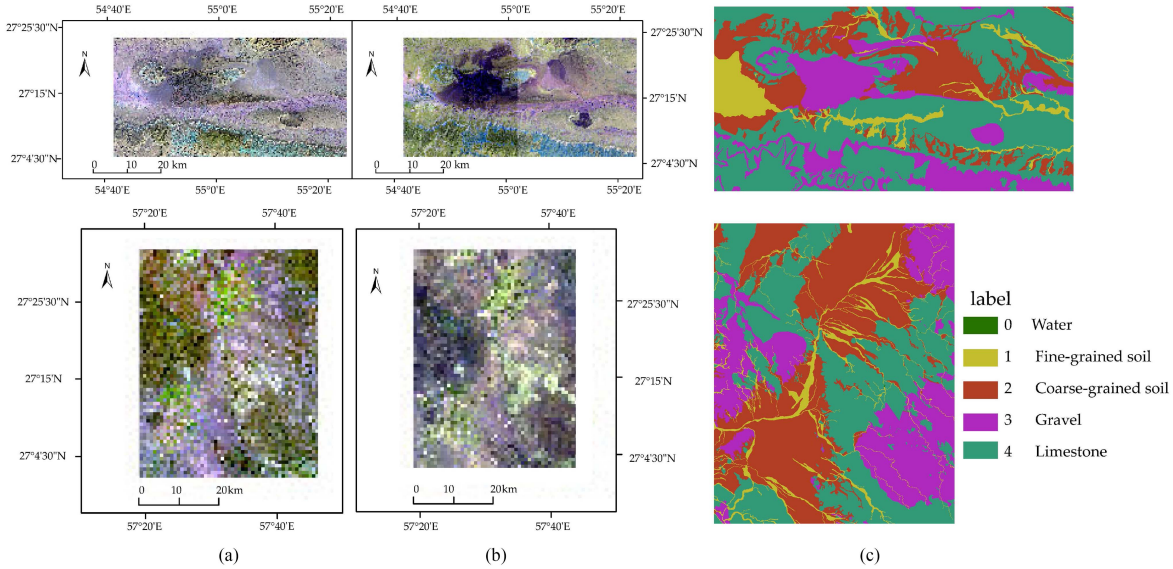

Fig. 3. Partial images from Landsat 8 and Sentinel-2 satellite remote sensing within the study area, along with their corresponding label maps. (a) Landsat8. (b) Sentinel-2. (c) Ground truth.

B. Multimodal Dataset

The complexity of the Earth's surface environment, coupled with the potential uncertainties in sensor imaging, affects the interpretation of rock, soil, and water features among geological elements. Therefore, in this study, Landsat 8 and Sentinel-2 satellite imagery data were utilized as the foundational data for the interpretation of these elements. However, it is important to note that the resolution of these satellite images inherently limits the clarity and detail that can be captured. Fig. 3 visually displays partial images from Landsat 8 and Sentinel-2 remote sensing satellites, along with their corresponding label maps, within the study area. Given these resolution constraints, Even experienced geological experts require field investigations to supplement their visual interpretations when dealing with such images. In addition, the integration of data from these different sensors, despite the resolution-induced blurriness, can be instrumental in validating the superiority of the proposed model.

Landsat 8: This satellite is equipped with a multispectral imaging system consisting of 11 bands, enabling the acquisition of medium-resolution images ranging from 15 to 100 m, covering the Earth's land surface and polar regions [43], [44]. In this study, a composite image was generated by merging bands 2, 3, 4, and 8 from the Landsat 8 dataset. The resulting fused image possesses a spatial resolution of 15 m. Subsequently, the dataset was carefully cropped to obtain 2562 distinct images depicting

TABLE I
PROPORTIONS OF THE FIVE CATEGORIES OF ROCK, SOIL, AND WATER FEATURES IN THE LANDSAT 8 DATASET

Data type	Pixel	Categories	Proportion (%)	others
Landsat8	0	Water	8.4	——
	1	Fine-grained soil	6.9	Soil particles with diameter d (mm) $\leq$ 0.075 mm and a proportion greater than 50%
	2	Coarse-grained soil	28.4	Soil particles with diameter 0.075 mm $< d$ (mm) ≤ 60
	3	Gravel	15.9	This is composed of more than 50% coarse fragments with a diameter greater than 2 mm
	4	Limestone	40.4	The predominant chemical composition is calcium carbonate, with the mineral component mainly composed of microcrystalline calcite

TABLE II
PROPORTIONS OF THE FIVE CATEGORIES OF ROCK, SOIL, AND WATER FEATURES IN THE SENTINEL-2 DATASET

Data type	Pixel	Categories	Proportion (%)	others
Sentinel-2	0	Water	7.9	——
	1	Fine-grained soil	6.9	Soil particles with diameter d (mm) $\leq$ 0.075 mm and a proportion greater than 50%
	2	Coarse-grained soil	28.3	Soil particles with diameter 0.075 mm $< d$ (mm) ≤ 60
	3	Gravel	16.2	This is composed of more than 50% coarse fragments with a diameter greater than 2 mm
	4	Limestone	40.7	The predominant chemical composition is calcium carbonate, with the mineral component mainly composed of microcrystalline calcite

rock, soil, and water features, each measuring 256×256 pixels. Notably, every image in the dataset is unique, without any duplications. Fig. 3(a) and (c) visually display Landsat 8 remote sensing images captured within the study area, accompanied by their corresponding labels. Table I provides a comprehensive breakdown of the proportions of the five categories of rock, soil, and water features (pixel value labels: $0-4$) found in the study area. It is evident from the table that the dataset exhibits an imbalanced distribution of rock, soil, and water features, representing a salient characteristic of the data.

Sentinel-2: This is a multispectral imaging instrument with 13 spectral bands, providing spatial resolutions ranging from 10 to 60 m [45]. In this study, the bands 12 (20-m resolution), 11 (20-m resolution), and 2 (10-m resolution) from Sentinel-2 were selected to create a composite image. The fused image has a resolution of 10 m. From this image, a dataset of 5745 unique rock, soil, and water images was extracted, each with a size of 256×256 pixels, ensuring there were no duplicates in the dataset. Fig. 3(b) and (c) displays the Sentinel-2 remote sensing images of the study area, along with their corresponding labels. Table II presents the proportion of the five categories of rock, soil, and water features (pixel value labels: $0-4$) in the study area. In addition, Table II also reveals the presence of an imbalanced distribution of rock, soil, and water, posing a challenge in the analysis. Notably, the Sentinel-2 data resulted in a higher resolution of the different proportions of rock, soil, and water compared to those presented in Table I.

IV. METHODOLOGY

A. Architecture of Network

The CEMFFNet model is inspired by the PSPNet network model. It consists of two main components: CEE and MFF. The CEE module serves as the backbone structure of the model, designed to extract abundant features for the interpretation of rock, soil, and water features. It establishes the connections between the elements and highlights their crucial feature information. In this study, the ResNet101 network improved with a Lambda layer was employed as the CEE [29], [46]. The MFF is used to combine features from different scales, integrating additional global information with the original features to enhance the network's discriminative ability for objects of different scales. Fig. 4 showcases the overall architecture of the network model and the data transmission method, primarily focusing on the following two aspects:

1) CEE: The CEE module incorporates the ResNet101 network and the Lambda layer. During the process of feature learning, DL networks aim to capture both global and salient features. ResNet101 is employed to the extract contextual features of geological elements, including rocks, soil, and water. The Lambda module transforms the context into a linear function to establish correlations between content and position features, significantly reducing memory consumption. Consequently, the introduction of the Lambda layer significantly enhances ResNet101's capability to extract the contextual features of geological elements, facilitating the extraction of multiple element features while establishing their intricate connections. This effectively resolves the challenge of distinguishing geological element features.
2) MFF: The MFF module integrates the SimAM attention mechanism and spatial pyramid pooling (SPP). The SimAM attention mechanism evaluates the feature information of each channel or spatial position and further explores the dependencies between feature channels and spatial positions. It adaptively learns the significance of

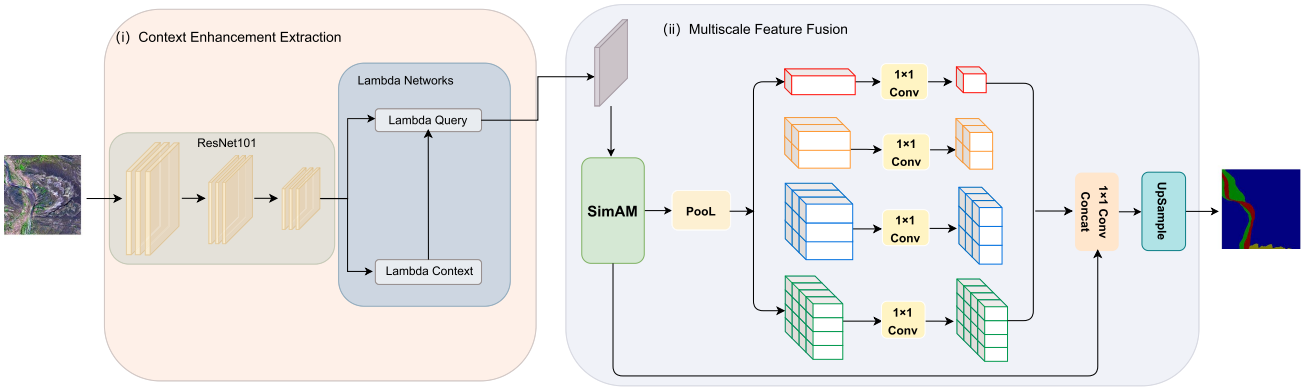

Fig. 4. Contextually enhanced multiscale feature fusion network (CEMFFNet).

each channel in the input feature map. By synergistically employing these two techniques, the module dynamically enhances the weights of informative features while suppressing irrelevant information. It seamlessly integrates the enriched contextual features from previous layers with multiple features at different scales, thereby amplifying the discriminative prowess of multiscale information within the features.

B. Context Enhancement Extraction

The ResNet101 network employs fixed convolutional kernels to extract feature information. However, this approach presents challenges in achieving an optimal receptive field and capturing comprehensive contextual feature information. Specifically, when dealing with geological elements, variations in features among different regions of the same rock type and similarities in features among different types of rock formations pose difficulties in precise interpretation. Consequently, accurately interpreting geological elements, including rocks, soil, and water, becomes a complex task. This challenge is prevalent in current DL methods applied to geological interpretation.

To capture the intricate and vital contextual feature information of rock, soil, and water features, this study incorporates the Lambda layer to augment the ResNet101 residual neural network, yielding the CEE module, as depicted in Fig. 5. The detailed steps are as follows:

1) Initially, the input data undergoes a $7 \times 7 \times 64$ convolution operation, followed by a 3×3 max pooling operation with a stride of 2. These steps effectively reduce the image size to a quarter of its original dimensions. Subsequently, the image is passed through three sets of Bottleneck modules within the ResNet101 network, each comprising three Bottleneck blocks, yielding an output feature map of size $7 \times 7 \times 1024$. The ResNet residual neural network implements skip connections, allowing input information to bypass and contribute to the final output. This innovative architecture facilitates lower level networks in concentrating exclusively on learning distinctive features from higher level networks, thereby

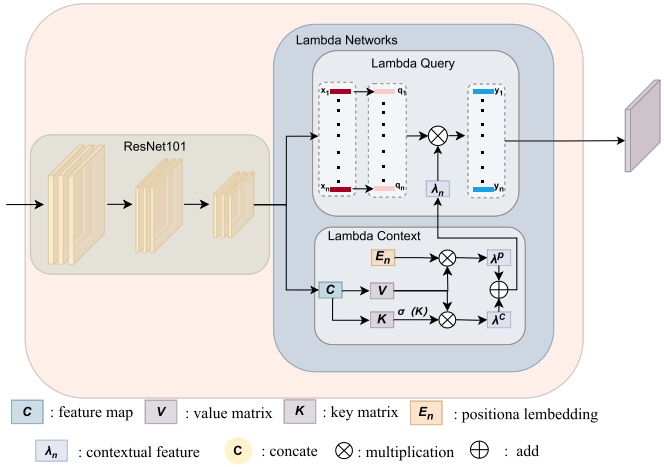

Fig. 5. CEE module schematic.

significantly upholding the integrity of information. Moreover, this design mitigates challenges like gradient vanishing or explosion during training and the potential degradation of multilayer network models as they learn parameters for nonlinear mapping.

2) The feature map C is subjected to linear projection to compute the key matrix K and value matrix V. Subsequently, Softmax normalization is applied to derive $\bar{K}$. The value matrix V and $\bar{K}$ contribute to the computation of the content Lambda function λ_n^c, while the value matrix V, along with the positional embedding relation E_n, is utilized to compute the positional Lambda function λ_n^p, where (n, m) represents the contextual position. The Lambda layer serves to model the correlation between feature content and spatial positions within an image. This layer transforms contextual information into a singular linear function, known as the Lambda function. In this research, the Lambda layer is employed to establish correlations among features and bolster the extraction of contextual feature information

$$K = CW_K \tag{1}$$

$$V = CW_V \tag{2}$$

$$\sigma(K) = \mathrm{Softmax}(K,\ \mathrm{axis}\ = m) \tag{3}$$

$$E_n \in R^{|m|\times|\boldsymbol{k}|} \tag{4}$$

$$\lambda^{\mathrm{c}} = \sum_{\mathrm{m}} \bar{K}_{\mathrm{m}} V_m^{\mathrm{T}} \tag{5}$$

$$\lambda_n^p = \sum_m E_{nm} V_m^T \tag{6}$$

where $|k|$ and $|\mathbf{v}|$ represent the query and value depths.

3) The contextual content λ_n^c and positional information λ_n^p are then fused to produce the contextual feature information λ_n. The input feature map X_n is employed to calculate the query item q_n through a learnable linear transformation. The contextual feature information λ_n dynamically interacts with the query item q_n, yielding the output result y_n. This process enables the modeling of interactions between content and position in global, local, or masked contexts, enabling more efficient computations

$$\lambda_n = \lambda^c + \lambda_n^p \in R^{|k|\times|v|} \tag{7}$$

$$y_n = \lambda_n^T q_n \in \mathbb{R}^{|v|}. \tag{8}$$

C. Multiscale Feature Fusion

To effectively integrate the comprehensive and critical feature maps of rock, soil, and water obtained from the CEE module, the CEMFFNet incorporates the MFF. MFF takes inspiration from the pyramid pooling module in the PSPNet network model to merge features, as depicted in Fig. 4(ii). The data input process of the MFF is as follows: The feature map M acquired from the CEE module is fed into the SimAM attention mechanism. The energy function e_t is employed to uncover the significant feature information within feature map M. The resulting features are then utilized as calibrated features, serving as inputs for the pyramid pooling operation. The pyramid pooling structure utilizes pooling layers of four different scales: 1×1, 2×2, 3×3, and 6×6, to extract global semantic information. This generates four sets of features at various scales, which are subsequently concatenated and fused with the calibrated features to obtain multiscale fusion features. The multiscale fusion feature map is further processed through a 1×1 convolutional layer and upsampling operations to yield the final interpreted image.

The essence of SimAM attention resides in the critical evaluation of individual neurons, facilitating superior attention implementation. Generally, neurons that are abundant in information exhibit distinctive discharge patterns compared to their neighboring counterparts. These active neurons possess spatial inhibitory capabilities, allowing them to suppress the activity of surrounding neurons. In visual signal processing, active neurons assume a greater significance and are endowed with heightened importance. The most straightforward and efficacious method of identifying pivotal neurons is to gauge the linear separability between the target neuron and its counterparts. Thus, an energy function is defined for each neuron as follows:

$$e_t\left(w_t, b_t, \mathrm{y}, x_i\right) = \left(y_t - \hat{t}\right)^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}\left(y_o - \hat{x}_i\right)^2 \tag{9}$$

$$\hat{t} = w_t t + b_t \tag{10}$$

$$\hat{x}_i = w_t x_i + b_t \tag{11}$$

$$M = H \times W \tag{12}$$

where $\hat{t}$ represents the target neuron and $\hat{x}_i$ denotes the remaining neurons. Equations (10) and (11) represent the linear transformations between $\hat{t}$ and $\hat{x}_i$. In these equations, w_t and b_t represent the weights and biases of the linear transformation, while i represents the spatial dimension index. M represents the total number of neurons in the respective channels.

To achieve linear separability between the target neuron t and other neurons within the same channel, minimizing e_t is typically pursued. When $y_t = \hat{t}$ and $\hat{x}_i = y_o$, e_t attains its minimum value. Using binary classification as an example, wherein $y_t = 1$ and $y_o = -1$, and incorporating a regularization term, the final form of the energy function is defined as follows:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2. \tag{13}$$

In theory, each channel consists of M energy functions, and solving them one by one entails a significant computational load. Utilizing (13), the following closed-form solution can be obtained:

$$w_t = -\frac{2(t-u_t)}{(t-u_t)^2 + 2\sigma_t^2 + 2\lambda} \tag{14}$$

$$b_t = -\frac{1}{2}(t+u_t)w_t \tag{15}$$

$$\mu_t = \frac{1}{M-1}\sum_{i=1}^{M-1} x_i \tag{16}$$

$$\sigma_t^2 = \frac{1}{M-1}\sum_{i=1}^{M-1}(x_i - \mu_t)^2. \tag{17}$$

Here, μ_t and σ_t^2 represent the mean and variance, respectively, of the pixels within the channel excluding the target neuron. In theory, all pixels within each channel share the same distribution, making the closed-form solution derived from a single channel applicable to any channel. Consequently, the following formula for the minimum energy is obtained:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda}. \tag{18}$$

In (18), a lower energy implies a greater disparity between neuron t and its neighboring neurons, indicating higher importance. Therefore, the significance of a neuron can be computed through e_t^*. As a result, by deriving the energy function, essential neurons can be uncovered. According to the definition of the attention mechanism, attention can bring about a gain effect and enhance

the processing of features. A scaling operation is applied to weight the features, resulting in :

$$\tilde{X} = \text{Sigmoid}\left(\frac{1}{E}\right) \odot X. \tag{19}$$

V. Experiments and Analysis

A. Experimental Settings

Section II introduces the dataset used in this study, which was divided into training, validation, and test sets using a random partitioning method with a ratio of $7:1:2$, respectively. The CEMFFNet model was evaluated on two satellite image datasets: Landsat 8 and Sentinel-2. In this experiment, a range of comparative networks were selected from different perspectives, including classical encoder–decoder structure networks, such as FCN [47] and UNet [48]. In addition, spatial pyramid structure networks, such as DeeplabV3+ [49] and PSPNet [50], were employed. Another category of models included real-time lightweight semantic segmentation models, namely BiSeNetV2 [51], LEDNet [52], and Segformer [53]. The experimental environment for these network models remained consistent across both datasets.

The study was carried out on a workstation featuring dual Intel XEON E5-2683V4 CPUs and dual NVIDIA RTX 2080Ti GPUs. The hyperparameters were configured as follows: a batch size of 16, a learning rate of 0.007, SGD was chosen as the optimizer, and the complete experimental procedure encompassed 200 epochs.

B. Evaluation Metrics

The accuracy of the experimental results was evaluated using the metrics oPA, IoU, and mIoU. The oPA and mIoU metrics were used to compare the performance of the different models, while the oPA and IoU were used to evaluate the interpretation results of the five categories of rock, soil, and water features.

Equation (20) represents the calculation process of oPA, where k denotes the number of element categories. p_{ij} represents the number of pixels that are actually of class i but are predicted as class j

$$oPA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}. \tag{20}$$

Equation (21) represents the IoU calculation process used to measure the ratio between the intersection and union of the sets of true values and predicted values

$$\text{IoU} = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{21}$$

Equation (22) represents the mIoU calculation process used to calculate the intersection over union ratio for each class and then take the average

$$\text{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{22}$$

C. Ablation Experiments

Ablation experiments were performed to validate the efficacy of the CEE module and the MFF module within the CEMFFNet model. Considering the inspiration from the PSPNet model, the PSPNet model was selected as the baseline model for this experiment [50]. To ensure the reliability of the experimental results, multiple trials were conducted, and the average values were calculated. In this study, the CEMFFNet model was tested on two distinct remote sensing satellite datasets, namely Landsat 8 and Sentinel-2, to assess the network model's robustness and ensure its dependable engineering performance.

Table III showcases the outcomes obtained from the ablation experiments. The PSPNet, serving as the baseline model, attained an oPA of 73.4% and an mIoU of 55.7% on the Landsat 8 dataset. In the Sentinel-2 dataset, the oPA stood at 72.3%, while the mIoU reached 56.9%. The integration of a solitary module led to notable enhancements in both oPA and mIoU. Notably, the inclusion of the CEE module had a more pronounced impact on the increase in oPA and mIoU in both datasets. In the Landsat 8 dataset, simultaneously incorporating both the CEE and MFF modules yielded an oPA of 80.5% and an mIoU of 65.7%. This corresponds to an approximate increase of 7.1% in oPA and 10% in mIoU compared to the baseline model. Similarly, in the Sentinel-2 dataset, the CEMFFNet model yielded oPA and mIoU of 80.4% and 66.3%, respectively, demonstrating improvements of around 8.1% in oPA and 9.4% in mIoU compared to the baseline model. The CEE module effectively captured intricate and crucial feature information pertaining to rock, soil, and water features, while the MFF module adeptly fused the acquired feature maps. The synergistic integration of these two modules substantially fortified the model's interpretive capabilities. Consequently, the results from these experiments provide compelling evidence of the exceptional performance of the CEMFFNet model in the interpretation of rock, soil, and water features.

D. Comparisons

The experimentation meticulously benchmarked against several emblematic DL semantic segmentation models to evaluate the interpretative acumen of the CEMFFNet model on the Landsat8 and Sentinel-2 datasets, while maintaining uniformity in model parameters and experimental ambiance. Table IV provides a lucid representation of the aggregate oPA and mIoU for the CEMFFNet model across the Landsat8 and Sentinel-2 datasets. From the tabulation, it is discernible that the oPA and mIoU metrics for both UNet and DeeplabV3+ marginally trail the model introduced in this manuscript, both underpinned by an encoder–decoder architecture. DeeplabV3+ leverages a spatial pyramid structure, bolstering the network's multiscale attributes and thereby enhancing segmentation precision. Springing from this, the CEMFFNet model championed in this study is architecturally grounded in the encoder–decoder framework augmented with pyramid features. When juxtaposed against the archetypal PSPNet model, the CEMFFNet introduced herein notably pares down the parameter load, ensuring an ascended precision while achieving reduced complexity.

TABLE III
ABLATION EXPERIMENT RESULTS ON LANDSAT 8 AND SENTINEL-2 DATASETS

Method Name	Landsat8	Sentinel-2	MFF	CEE	oPA(%)	mIoU(%)
PSPNet	√				73.4	55.7
		√			72.3	56.9
Ours(CEMFFNet)	√		√		73.7	57
	√			√	78.2	64.8
	√		√	√	**80.5**	**65.7**
		√	√		73.6	57.9
		√		√	78.6	65.6
		√	√	√	**80.4**	**66.3**

* The best results are highlighted in color, with green for the Sentinel-2 dataset and blue for the Landsat 8 dataset.

TABLE IV
INTERPRETATION PERFORMANCE AND PARAMETER COUNT OF DIFFERENT MODELS ON THE LANDSAT 8 AND SENTINEL-2 DATASETS

Method Name	Data source	oPA(%)	mIOU(%)	Parameters(M)
FCN [47]	Landsat8 Sentinel-2	69.6 68.4	53.4 51.6	134.81
Unet [48]	Landsat8 Sentinel-2	75.2 71.3	59.5 58.2	29.20
DeeplabV3+ [49]	Landsat8 Sentinel-2	78.1 76.9	62.9 63.1	37.05
PSPNET [50]	Landsat8 Sentinel-2	73.4 72.3	55.7 56.9	25.70
BiSeNetV2 [51]	Landsat8 Sentinel-2	70.6 67.9	52.6 52.3	2.93
LEDNet [52]	Landsat8 Sentinel-2	69.7 69.6	52.7 54.6	**2.35**
Segformer [53]	Landsat8 Sentinel-2	64.8 50.5	43.8 38.5	31.01
MANet [54]	Landsat8 Sentinel-2	68.9 69.2	52.5 53.8	5.3
Mask2Former [55]	Landsat8 Sentinel-2	75.8 75.3	61.3 62.1	44.23
Ours(CEMFFNet)	Landsat8 Sentinel-2	**80.2** **80.4**	**65.2** **66.3**	45.01

* The best results are highlighted in color, with green for the Sentinel-2 dataset and blue for the Landsat 8 dataset.

Distinct from the Landsat8 dataset, the Sentinel-2 dataset boasts a spatial resolution of 10 m, markedly surpassing that of its Landsat8 counterpart. Such an augmentation in resolution endows remote sensing imagery with heightened detail. As a result, the CEMFFNet model delineated in this discourse exhibits a notable enhancement in oPA and mIoU by 0.2% and 1.1%, respectively, when juxtaposed against the Landsat8 dataset. This observation cogently underscores the significance of amplifying image resolution as a strategy to bolster interpretative precision. Yet, other models intriguingly fail to echo this attribute. This discrepancy may stem from their ambivalence in extracting salient contextual attributes associated with rock, soil, and water, indicative of either an inability to discern the rich classification nuances of these constituents or an oversight in accentuating seminal feature details.

Both Landsat8 and Sentinel-2 datasets classify into five geological categories: Water, fine-grained soil, coarse-grained soil, Gravel, and Limestone. Tables V and VI, respectively, delineate the performance metrics, namely PA and IoU, for various models across these categories in both datasets. This article introduces the CEMFFNet model, which exhibits superior classification prowess, as indicated by its PA and IoU, for all five categories in the Landsat8 dataset. Moreover, within the Sentinel-2 dataset, the CEMFFNet model distinctly excels in the interpretation of fine-grained soil, coarse-grained soil, Gravel, and Limestone.

Due to the uneven distribution of rock, soil, and water elements, there is an inherent challenge: the internal characteristics of an element might greatly differ, while interelement characteristics might bear minimal discrepancies. For instance, both fine-grained soil and Gravel are categorized based on particle proportion. Specifically, fine-grained soil possesses less than 50% coarse particles, while Gravel is composed of over 50% fragments with a diameter exceeding 2 mm. Such classification paradigms inevitably result in the aforementioned discrepancies. In addition, fine-grained soil represents a relatively minor proportion of all samples, at a mere 6.9%, leading to consistently

TABLE V
OPA OF DIFFERENT NETWORK MODELS FOR ROCK, SOIL, AND WATER FEATURES ON THE LANDSAT 8 AND SENTINEL-2 DATASETS (%)

Method Name	Data source	Water	Fine-grained soil	Coarse-grained soil	Gravel	Limestone
FCN	Landsat8	94.1	38.3	79.8	36.0	76.9
	Sentinel-2	**98.3**	35.5	73.8	21.7	82.8
Unet	Landsat8	96.9	49.5	78.4	48.7	82.6
	Sentinel-2	98.0	60.3	77.3	51.6	72.2
DeeplabV3+	Landsat8	**97.3**	48.5	84.4	59.9	81.0
	Sentinel-2	96.6	57.3	85.3	54.0	80.3
PSPNET	Landsat8	95.3	39.0	81.2	43.8	80.0
	Sentinel-2	95.9	55.6	73.1	44.1	77.9
BiSeNetV2	Landsat8	94.7	34.9	79.8	32.5	79.5
	Sentinel-2	96.3	41.8	75.1	26.5	78.9
LEDNet	Landsat8	94.7	37.9	80.3	35.6	74.8
	Sentinel-2	94.5	46.2	82.0	39.0	73.2
Segformer	Landsat8	89.3	25.2	73.3	8.3	80.4
	Sentinel-2	97.3	28.7	29.5	15.7	73.4
MANet	Landsat8	93.6	37.5	79.8	35.8	74.9
	Sentinel-2	94.2	45.8	82.1	36.7	74.6
Mask2Former	Landsat8	95.8	46.9	83.1	60.1	79.8
	Sentinel-2	93.9	56.8	83.9	54.2	80.5
Ours(CEMFFNet)	Landsat8	96.5	**53.1**	**84.8**	**62.9**	**84.2**
	Sentinel-2	97.7	**60.9**	**85.5**	**57.0**	**83.4**

* The best results are highlighted in color, with green for the Sentinel-2 dataset and blue for the Landsat 8 dataset.

TABLE VI
IOU OF DIFFERENT NETWORK MODELS FOR ROCK, SOIL, AND WATER FEATURES ON THE LANDSAT 8 AND SENTINEL-2 DATASETS (%)

Method Name	Data source	Water	Fine-grained soil	Coarse-grained soil	Gravel	Limestone
FCN	Landsat8	92.8	30.6	56.8	27.6	59.2
	Sentinel-2	**96.4**	30.9	54.4	18.8	57.3
Unet	Landsat8	94.5	37.9	63.2	36.1	65.8
	Sentinel-2	95.8	41.4	61.5	33.5	58.9
DeeplabV3+	Landsat8	93.9	40.6	68.0	44.3	67.4
	Sentinel-2	94.3	46.9	66.1	41.4	67.0
PSPNET	Landsat8	89.5	30.6	61.7	33.4	62.9
	Sentinel-2	93.5	44.6	54.7	31.8	59.7
BiSeNetV2	Landsat8	91.6	27.4	58.1	25.5	60.6
	Sentinel-2	93.7	34.8	55.1	20.7	57.2
LEDNet	Landsat8	90.9	30.5	59.2	25.4	57.5
	Sentinel-2	91.6	36.8	58.7	28.0	57.9
Segformer	Landsat8	85.8	21.6	49.5	7.10	55.1
	Sentinel-2	93.0	24.7	20.1	12.5	42.2
MANet	Landsat8	90.3	32.2	58.7	25.9	55.8
	Sentinel-2	92.3	37.4	56.9	29.7	57.1
Mask2Former	Landsat8	91.5	38.3	67.4	44.9	68.2
	Sentinel-2	93.8	44.7	68.3	43.6	66.3
Ours(CEMFFNet)	Landsat8	**95.1**	**40.6**	**69.4**	**49.8**	**71.2**
	Sentinel-2	95.4	**50.1**	**67.7**	**45.6**	**70.2**

* The best results are highlighted in color, with green for the Sentinel-2 dataset and blue for the Landsat 8 dataset.

low interpretative accuracy across various network models. Among all models, water, coarse-grained soil, and Limestone categories consistently achieve the highest interpretative precision.

The prevailing models frequently surpass the CEMFFNet in delineating water accuracy. Yet, for the subsequent four geological constituents, their precision pales in comparison to CEMFFNet. A juxtaposition of imaging traits between water and the quartet of geological elements reveals water's color characteristics to be conspicuously pronounced and unequivocal relative to rock and soil. Such observations suggest that conventional DL paradigms demonstrate commendable resilience when interpreting overt and manifest land features, but their efficacy diminishes for geological facets characterized by variegated color nuances and limited discernibility. Consequently, sophisticated geological components demand more refined model configurations and feature distillation methodologies to bolster interpretation's accuracy and resilience. This insight provides an instrumental roadmap for ensuing endeavors in the interpretative studies of rock, soil, and water facets.

TABLE VII
CONFUSION MATRIX OF THE CEMFFNET MODEL ON THE LANDSAT 8 DATASET (VALIDATION DATA)

	Water	Fine-grained soil	Coarse-grained soil	Gravel	Limestone
Water	**1279532(97.1%)**	23988	8355	2165	3820
Fine-grained soil	42180	**737090(53.9%)**	476621	25219	86628
Coarse-grained soil	669	136699	**4621513(87.7%)**	194026	315870
Gravel	910	17364	236741	**1801567(60.5%)**	918807
Limestone	298	34889	326840	658077	**4827348(82.6%)**

* Number of pixels where the true values and predicted values are equal.

TABLE VIII
CONFUSION MATRIX OF THE CEMFFNET MODEL ON THE LANDSAT 8 DATASET (TEST DATA)

	Water	Fine-grained soil	Coarse-grained soil	Gravel	Limestone
Water	**3217787(96.2%)**	97072	20704	475	7498
Fine-grained soil	34287	**1193852(53.9%)**	759094	76042	151886
Coarse-grained soil	8119	343613	**8227476(84.4%)**	349322	772810
Gravel	4070	95693	432331	**3410152(63.4%)**	1438786
Limestone	722	84500	847283	1059981	**11051949(84.7%)**

* Number of pixels where the true values and predicted values are equal.

TABLE IX
CONFUSION MATRIX OF THE CEMFFNET MODEL ON THE SENTINEL-2 DATASET (VALIDATION DATA)

	Water	Fine-grained soil	Coarse-grained soil	Gravel	Limestone
Water	**3182593(96.2%)**	102349	13410	2713	8678
Fine-grained soil	66031	**1333418(52.3%)**	763895	68020	318353
Coarse-grained soil	2827	211332	**8755633(84.4%)**	357365	1052579
Gravel	6950	39036	456644	**3907047(62.9%)**	1802789
Limestone	77	50593	849038	1145609	**13120685(86.5%)**

* Number of pixels where the true values and predicted values are equal.

TABLE X
CONFUSION MATRIX OF THE CEMFFNET MODEL ON THE SENTINEL-2 DATASET (TEST DATA)

	Water	Fine-grained soil	Coarse-grained soil	Gravel	Limestone
Water	**6335089(97.8%)**	105441	23784	1357	8771
Fine-grained soil	88899	**2684710(52.7%)**	1641042	161728	518529
Coarse-grained soil	11949	506310	**17478351(83.9%)**	630125	2213965
Gravel	8374	74016	1051016	**7488264(60.2%)**	3822269
Limestone	1719	118052	1911105	1931692	**26746451(87.1%)**

* Number of pixels where the true values and predicted values are equal.

It is significant to observe that, in the interpretation of fine-grained soil, the majority of models display a pronounced superiority in accuracy using Sentinel-2 over Landsat8. Due to this distinction, the CEMFFNet model exhibits a higher overall interpretative accuracy with the Sentinel-2 dataset than with the Landsat8 dataset. Meanwhile, the interpretative precision for the remaining four geological elements remains relatively aligned between both datasets.

In contrast to the PSPNet benchmark model, the model introduced in this study achieves the utmost level of interpretive precision within all five categories. The experimental outcomes clarify that the incorporated FEE module and MFF module in this research make substantive contributions toward unearthing contextual feature information of geological elements. This significant contribution is substantiated thoroughly across both datasets, firmly establishing the supremacy of the CEMFFNet model.

Tables VII and VIII present the confusion matrices illustrating the performance of the CEMFFNet model on the validation data and test data from the Landsat8 dataset, respectively. In these tables, the rows and columns correspondingly represent the predicted label and the true label. In addition, Tables IX and X exhibit the confusion matrices for the CEMFFNet model's performance on the validation data and test data from the Sentinel-2 dataset. The insights gleaned from these tables underscore that the proposed CEMFFNet model exhibits elevated accuracy in

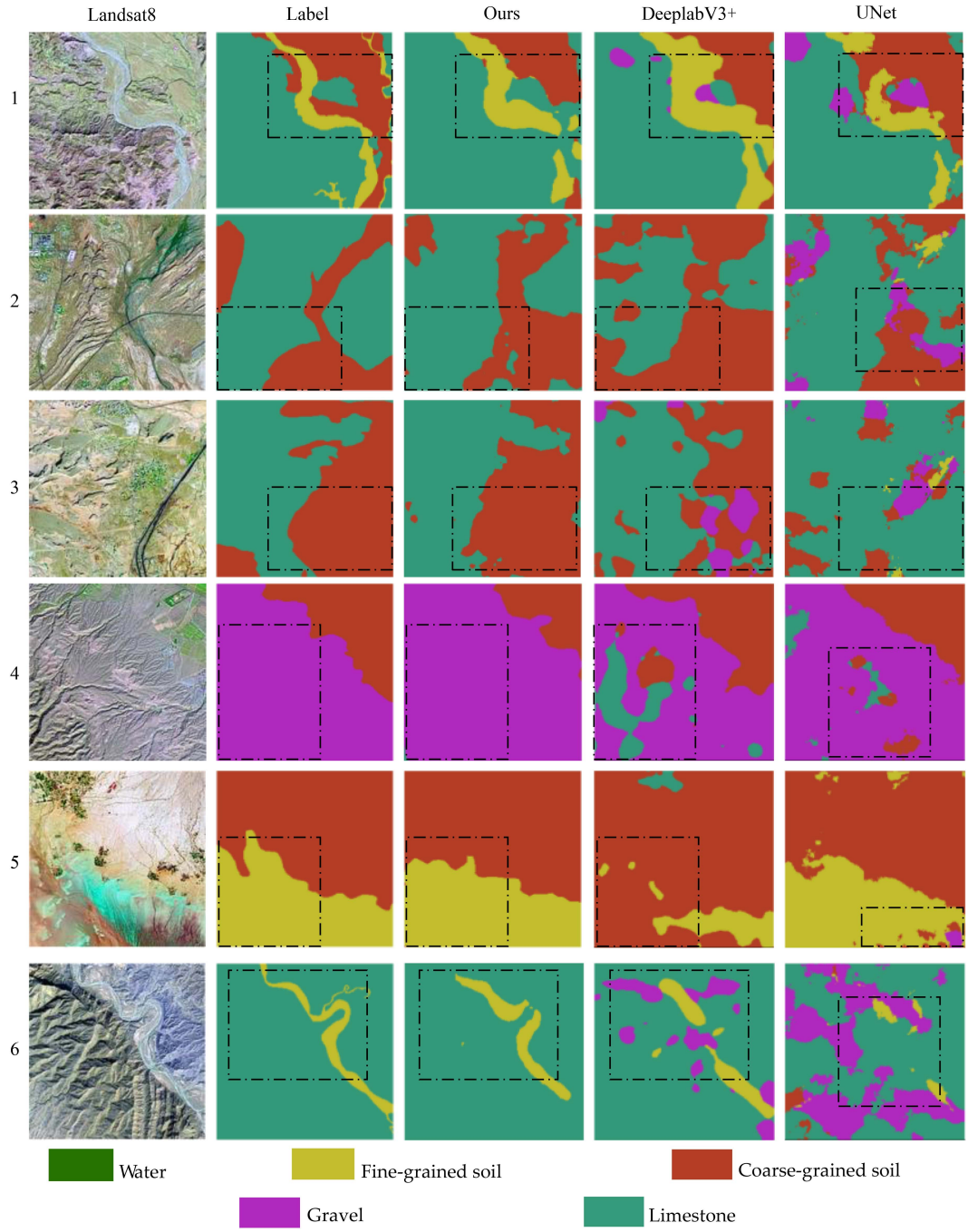

Fig. 6. Part of the predicted examples by CEMFFNet, DeeplabV3+, and UNet on the Landsat 8 dataset.

classifying water, coarse-grained soil, and limestone. However, there remains a scope for further refining the accuracy concerning the interpretation of fine-grained soil and gravel. This observation stems from the fact that fine-grained soil is prone to being inaccurately classified as coarse-grained soil, while gravel is susceptible to erroneous classification as limestone. Such misclassifications impede the overall enhancement of interpretive precision.

Fig. 6 visually showcases the predictive prowess of the three notably adept models, namely CEMFFNet, DeeplabV3+, and Unet, across the Landsat8 dataset. Similarly, Fig. 7 vividly presents the predictive outputs of these three models over the Sentinel-2 dataset. While this study's model exhibits occasional misinterpretations in specific regions, its predictive outputs remain characterized by minimal instances of misclassification. The visual representation elucidates that the predictions emanating from the CEMFFNet model bear a striking semblance to the verifiable ground truth labels, epitomizing the utmost parity. Conversely, the prognostications of DeeplabV3+ and UNet exhibit a comparatively higher incidence of misclassifications, thus contributing to a more pronounced distortion in their classification outcomes. The introduced

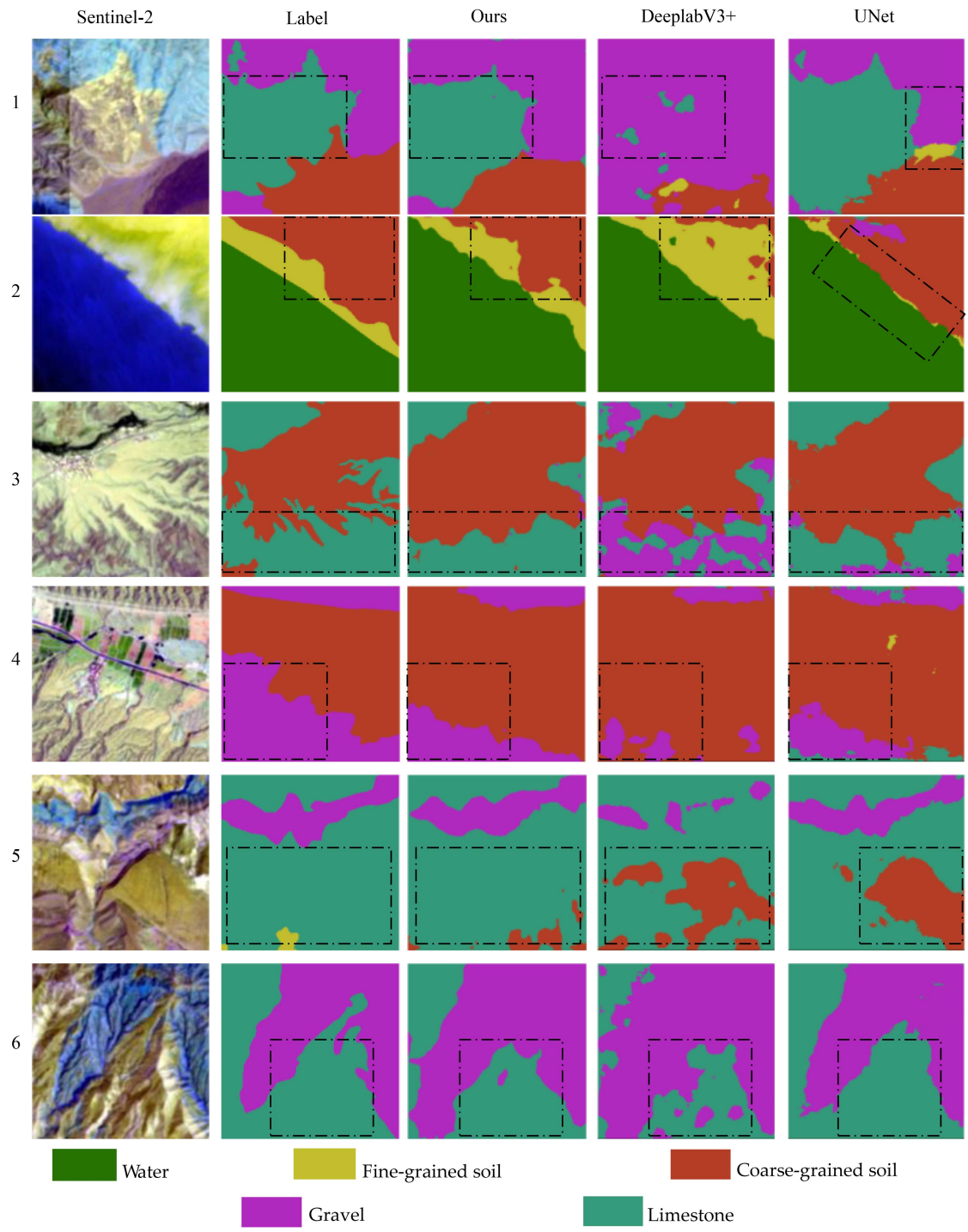

Fig. 7. Part of the predicted examples by CEMFFNet, DeeplabV3+, and UNet on the Sentinel-2 dataset.

CEMFFNet model impeccably demarcates the boundaries between distinct elemental categories, adeptly mitigating the quandary arising from the inherent interclass resemblance and intraclass variability between rock and soil classifications. This culminates in the CEMFFNet model effectively achieving the desired outcome of intraclass cohesion and interclass differentiation.

To demonstrate the efficacy of the CEMFFNet model in practical applications, this study presents the predictive outcomes for a substantial geographical area, as illustrated in Fig. 8. The remote sensing imagery employed originates from Landsat 8 and corresponds to the initial image depicted in Fig. 3. It is pertinent to note that the training dataset for this study was compiled through a randomized sampling approach. The displayed result encompasses both the training and testing datasets, offering a comprehensive prediction based on a blended dataset. Consequently, the overall pixel accuracy stands at 89.9%. While minor discrepancies are evident in some details of the prediction, the

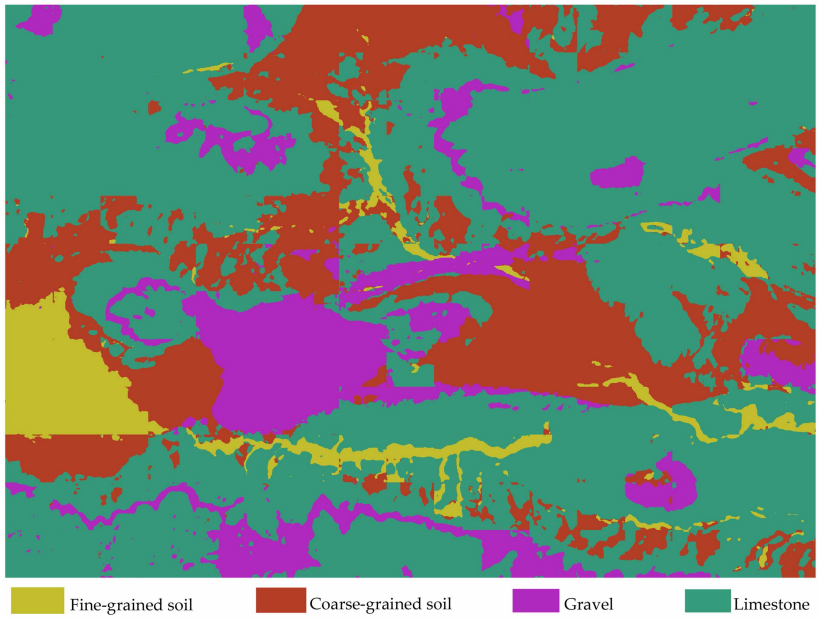

Fig. 8. Large-scale predictive mapping of Landsat 8 satellite remote sensing within the study area.

TABLE XI
PARAMETERS AND COMPUTATIONAL COMPLEXITIES OF PSPNET AND CEMFFNET

MODEL	INPUT SIZE	PARAMETERS	GFLOPs
PSPNet	$3 \times 256 \times 256$	65.70 (M)	19.69
CEMFFNet	$3 \times 256 \times 256$	45.01 (M)	44.87

fundamental structure and overall contour align closely with the actual labels, attesting to the CEMFFNet model's robustness and accuracy in large-scale dataset evaluations.

E. Complexity Analysis

To evaluate the computational overhead, the number of model parameters and computational complexity (GFLOPs) were selected as metrics to analyze the complexity of the DeepLab v3+ and the CEMFFNet model. The findings are delineated in Table XI. The experimental results showed that since the size of the Landsat 8 and Sentinel-2 images are identical, being 256×256, the input size does not affect the GFLOPs for both models. The parameter count of the PSPNet model is 1.5 times that of the CEMFFNet model, indicating that the addition of the FEE has a minimal impact on the parameter count of the CEMFFNet model. However, the PSPNet model has lower GFLOPs compared to the CEMFFNet model, indicating that the FEE significantly impacts computational efficiency. Overall, the computational overhead of the CEMFFNet model is within a reasonable range, considering the performance improvements it provides.

VI. CONCLUSION

To tackle the inherent challenges arising from the susceptibility of rock, soil, and water features to environmental cover and geological modifications, resulting in limitations in the observability of satellite remote sensing data, concealed geological features, complex ground feature distributions, and notable spatial variability of geological elements, this study introduces the CEMFFNet model as a solution. CEMFFNet comprises two essential components: the CEE and the MFE. The former enhances the contextual correlation among rock, soil, and water features, effectively capturing intricate and critical feature information associated with these elements. Meanwhile, the latter integrates multiscale feature information on rock, soil, and water features. The study area is situated in the Iranian region, renowned for its rich variety of rock formations, soil types, and water bodies. To encompass the breadth of these geological element categories, Landsat 8 and Sentinel-2 datasets were meticulously assembled for this specific region. The experimental findings unequivocally demonstrate that the CEMFFNet model surpasses other comparative models in terms of overall interpretive accuracy across both datasets. This underscores the remarkable potential of the CEMFFNet model in the interpretation of rock, soil, and water features.

The interpretation of rock, soil, and water elements presents a complex challenge in the realm of geological environmental remote sensing. This study provides an in-depth analysis of the obstacles posed by real-world geological environmental scenarios in interpreting these elemental constituents. It identifies two key issues: the coverage of geographical environments and the modifications resulting from geological activities both impact the interpretation of rock, soil, and water elements. This aspect has not received adequate attention in prior research on the intelligent interpretation of these geological constituents. The findings of this study can offer fresh insights to researchers engaged in regional geological mapping and surveying. However, there is room for improvement in our work, particularly in accurately interpreting fine-grained soil and gravel elements,

where results show lower precision. Uneven distribution of these elemental categories within the dataset contributes to this disparity. Moreover, our network may have limitations in addressing the challenge of class imbalance. Hence, providing targeted guidance for the precise classification of fine-grained soil and gravel is crucial to enhancing the overall interpretive accuracy of the model. In future endeavors, comprehensive investigations will be conducted to address this issue and further enhance the model's performance in these specific geological element categories.

REFERENCES

[1] M. J. Cracknell and A. M. Reading, "Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information," *Comput. Geosci.*, vol. 63, pp. 22–33, 2014.

[2] Z. Wang, R. Zuo, and F. Yang, "Geological mapping using direct sampling and a convolutional neural network based on geochemical survey data," *Math. Geosci.*, vol. 55, pp. 1035–1058, 2023.

[3] K. He, Y. Dong, W. Han, and Z. Zhang, "An assessment on the off-road trafficability using a quantitative rule method with geographical and geological data," *Comput. Geosci.*, vol. 177, 2023, Art. no. 105355.

[4] T. Yamaguchi et al., "Hayabusa2-Ryugu proximity operation planning and landing site selection," *Acta Astronautica*, vol. 151, pp. 217–227, 2018.

[5] M. Zhu et al., "Strengthening mechanism of granulated blast-furnace slag on the uniaxial compressive strength of modified magnesium slag-based cemented backfilling material," *Process Saf. Environ. Protection*, vol. 174, pp. 722–733, 2023.

[6] R. Fan, L. Wang, J. Yan, W. Song, Y. Zhu, and X. Chen, "Deep learning-based named entity recognition and knowledge graph construction for geological hazards," *ISPRS Int. J. Geo- Inf.*, vol. 9, no. 1, 2019, Art. no. 15.

[7] S. Fathololoumi, A. R. Vaezi, S. K. Alavipanah, A. Ghorbani, D. Saurette, and A. Biswas, "Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: A case study in Iran," *Sci. Total Environ.*, vol. 721, 2020, Art. no. 137703.

[8] M. A. Abdelkader, Y. Watanabe, A. Shebl, H. A. El-Dokouny, M. Dawoud, and Á. Csámer, "Effective delineation of rare metal-bearing granites from remote sensing data using machine learning methods: A case study from the Umm Naggat area, Central Eastern Desert, Egypt," *Ore Geol. Rev.*, vol. 150, 2022, Art. no. 105184.

[9] J. Lu et al., "Lithology classification in semi-arid area combining multi-source remote sensing images using support vector machine optimized by improved particle swarm algorithm," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 119, 2023, Art. no. 103318.

[10] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, pp. 1082–1086, 2019.

[11] C. Wu, X. Li, W. Chen, and X. Li, "A review of geological applications of high-spatial-resolution remote sensing data," *J. Circuits, Syst. Comput.*, vol. 29, no. 06, 2020, Art. no. 2030006.

[12] X. Zhang, Y. Zhou, and J. Luo, "Deep learning for processing and analysis of remote sensing Big Data: A technical review," *Big Earth Data*, vol. 6, no. 4, pp. 527–560, 2022.

[13] R. Fan, J. Li, F. Li, W. Han, and L. Wang, "Multilevel spatial-channel feature fusion network for urban village classification by fusing satellite and streetview images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630813.

[14] R. Fan, J. Li, W. Song, W. Han, J. Yan, and L. Wang, "Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 111, 2022, Art. no. 102831.

[15] R. Fan, F. Li, W. Han, J. Yan, J. Li, and L. Wang, "Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630316.

[16] R. Fan, R. Feng, L. Wang, J. Yan, and X. Zhang, "Semi-MCNN: A semisupervised multi-CNN ensemble learning method for urban land cover classification using submeter HRRS images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4973–4987, 2020.

[17] Z. Benbahria, I. Sebari, H. Hajji, and M. F. Smiej, "Intelligent mapping of irrigated areas from landsat 8 images using transfer learning," *Int. J. Eng. Geosci.*, vol. 6, no. 1, pp. 40–50, 2021.

[18] Y. Li, W. Xu, H. Chen, J. Jiang, and X. Li, "A novel framework based on mask R-CNN and histogram thresholding for scalable segmentation of new and old rural buildings," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1070.

[19] R. Prabhu, B. Parvathavarthini, and A. R. Alaguraja, "Integration of deep convolutional neural networks and mathematical morphology-based postclassification framework for urban slum mapping," *J. Appl. Remote Sens.*, vol. 15, no. 1, 2021, Art. no. 014515.

[20] C. Persello and M. Kuffer, "Towards uncovering socio-economic inequalities using VHR satellite images and deep learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 3747–3750.

[21] M. Dixit, K. Chaurasia, and V. K. Mishra, "Automatic building extraction from high-resolution satellite images using deep learning techniques," in *Proc. Int. Conf. Paradigms Comput., Commun. Data Sci.*, 2021, pp. 773–783.

[22] G. Zhou, J. Xu, W. Chen, X. Li, J. Li, and L. Wang, "Deep feature enhancement method for land cover with irregular and sparse spatial distribution features: A case study on open-pit mining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401220.

[23] Y. Lu et al., "Remote sensing interpretation for soil elements using adaptive feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4505515.

[24] J. Lu, L. Han, X. Zha, and L. Li, "Lithology classification in semi-arid areas based on vegetation suppression integrating microwave and optical remote sensing images: Duolun county, inner Mongolia autonomous region, China," *Geocarto Int.*, vol. 37, no. 27, pp. 17044–17067, 2022.

[25] H. Dang, T. Zhang, Y. Li, G. Li, L. Zhuang, and X. Pu, "Population evolution, genetic diversity and structure of the medicinal legume, Glycyrrhiza Uralensis and the effects of geographical distribution on leaves nutrient elements and photosynthesis," *Front. Plant Sci.*, vol. 12, 2022, Art. no. 708709.

[26] Y. Wang, S. Li, Y. Lin, and M. Wang, "Lightweight deep neural network method for water body extraction from high-resolution remote sensing images with multisensors," *Sensors*, vol. 21, no. 21, 2021, Art. no. 7397.

[27] T. Que, Y. Wu, S. Hu, J. Cai, N. Jiang, and H. Xing, "Factors influencing public participation in community disaster mitigation activities: A comparison of model and nonmodel disaster mitigation communities," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, 2022, Art. no. 12278.

[28] M. Zhang et al., "A habitable earth and carbon neutrality: Mission and challenges facing resources and the environment in China–An overview," *Int. J. Environ. Res. Public Health*, vol. 20, no. 2, 2023, Art. no. 1045.

[29] I. Bello, "LambdaNetworks: Modeling long-range interactions without attention," in *Proc. Int. Conf. Learn. Representations*, 2021.

[30] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.

[31] I. Bachri, M. Hakdaoui, M. Raji, A. C. Teodoro, and A. Benbouziane, "Machine learning algorithms for automatic lithological mapping using remote sensing data: A case study from Souk Arbaa Sahel, Sidi Ifni Inlier, western anti-atlas, Morocco," *ISPRS Int. J. Geo- Inf.*, vol. 8, no. 6, 2019, Art. no. 248.

[32] T. S. Bressan, M. K. de Souza, T. J. Girelli, and F. C. Junior, "Evaluation of machine learning methods for lithology classification using geophysical data," *Comput. Geosci.*, vol. 139, 2020, Art. no. 104475.

[33] C. Kumar, S. Chatterjee, T. Oommen, and A. Guha, "Automated lithological mapping by integrating spectral enhancement techniques and machine learning algorithms using aviris-Ng hyperspectral data in gold-bearing granite-greenstone rocks in Hutti, India," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 86, 2020, Art. no. 102006.

[34] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–34, 2022.

[35] Y. Wang, X. Chen, and L. Wang, "Cyber-physical oil spill monitoring and detection for offshore petroleum risk management service," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 4586.

[36] Y. Wang, L. Wang, X. Chen, and D. Liang, "Offshore petroleum leaking source detection method from remote sensing data via deep reinforcement learning with knowledge transfer," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5826–5840, 2022.

[37] Z. Wang, R. Zuo, and H. Liu, "Lithological mapping based on fully convolutional network and multi-source geological data," *Remote Sens.*, vol. 13, no. 23, 2021, Art. no. 4860.

[38] H. Shirmard et al., "A comparative study of convolutional neural networks and conventional machine learning models for lithological mapping using remote sensing data," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 819.
[39] W. Han et al., "Geological remote sensing interpretation using deep learning feature and an adaptive multisource data fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4510314.
[40] X. Wang, R. Zuo, and Z. Wang, "Lithological mapping using a convolutional neural network based on stream sediment geochemical survey data," *Natural Resour. Res.*, vol. 31, no. 5, pp. 2397–2412, 2022.
[41] S. Doulabian, A. S. Toosi, G. H. Calbimonte, E. G. Tousi, and S. Alaghmand, "Projected climate change impacts on soil erosion over Iran," *J. Hydrol.*, vol. 598, 2021, Art. no. 126432.
[42] M. Farpoor, M. Neyestani, M. Eghbal, and I. E. Borujeni, "Soil–geomorphology relationships in Sirjan playa, south central Iran," *Geomorphology*, vol. 138, no. 1, pp. 223–230, 2012.
[43] K. W. Abdelmalik, "Landsat 8: Utilizing sensitive response bands concept for image processing and mapping of basalts," *Egyptian J. Remote Sens. Space Sci.*, vol. 23, no. 3, pp. 263–274, 2020.
[44] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2291.
[45] M. Marshall, M. Belgiu, M. Boschetti, M. Pepe, A. Stein, and A. Nelson, "Field-level crop yield estimation with PRISMA and Sentinel-2," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 191–210, 2022.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.
[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[51] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.
[52] Y. Wang et al., "Lednet: A lightweight encoder-decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1860–1864.
[53] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 12077–12090.
[54] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.
[55] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1290–1299.

Kang He received the master's degree in engineering in mining engineering from Central South University, Changsha, China, in 2021. He is currently working toward the Ph.D. degree in military geology with the China University of Geosciences, Wuhan, China.

His research interests include geological environmental remote sensing and off-road mobility analysis.

Zhijun Zhang received the master's degree in engineering in earth exploration and information technology from China University of Mining and Technology, Beijing, China, in 2012. He is currently working toward the Ph.D. degree in military geology with the China University of Geosciences, Wuhan, China.

His research interests is the application of remote sensing technology in geological survey.

Yusen Dong received the B.S. degree in geology, the M.S. degree in geography, and the Ph.D. degree in cartography and geographic information systems from the China University of Geosciences, Wuhan, China, in 1999, 2002, and 2007, respectively.

From 2006 to 2008, he was engaged in cooperative research in Interferometric Synthetic Aperture Radar, Differential InSAR, Permanent Scatterer InSAR, University of New South Wales, Sydney, Australia. He is currently an Associate Professor with the School of Computer, China University of Geosciences. His research interests include DInSAR and ground deformation monitoring, geological remote sensing and remote-sensing investigation of land and resources, and glacier change monitoring.

Depan Cai received the master's degree in engineering in computer technology from the China University of Geosciences, Wuhan, China, in 2023.

He is currently working with the Hubei Provincial Meteorological Service Center, Hubei Provincial Meteorological Service Desk. His research interests include remote sensing image processing and traffic weather analysis.

Yue Lu received the B.E. degree in chemical engineering and technology from Nanjing Forestry University, Nanjing, China, in 2020. He is currently working toward the M.E. degree in computer technology with the China University of Geosciences, Wuhan, China.

His research interests include deep learning, remote-sensing geological interpretation, and multimodal fusion.

Wei Han received the B.S., M.S., and Ph.D. degrees in geoscience information engineering from the China University of Geosciences, Wuhan, China.

He is an Associate Professor with the School of Computer Science, China University of Geosciences, Wuhan, China. His research interests include data management, high-performance computing, and high-resolution remote sensing image processing.