Received 20 March 2024, accepted 9 April 2024, date of publication 15 April 2024, date of current version 24 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3388564

RESEARCH ARTICLE

Typical Ground Object Recognition in Desert Areas Based on DYDCNet: A Case Study in the Circum-Tarim Region, Xinjiang, China

JUNFU FAN[1,2], YU GAO[1], ZONGWEN SHI[1], PING LI[1], AND GUANGWEI SUN[1]

[1]School of Civil Engineering and Geomatics, Shandong University of Technology, Zibo, Shandong 255000, China
[2]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Corresponding author: Guangwei Sun (sgw_sdut@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 42171413, and in part by the State Key Laboratory of Resources and Environmental Information System.

ABSTRACT Automatic feature semantic segmentation of remote sensing images is an extremely critical research direction in the field of geographic information science. Especially in the vast and complex desert area, the wide spatial distribution of surface features, complex feature texture characteristics and uneven sample classification bring great challenges to the recognition and segmentation of features. In response to the question, we propose an innovative semantic segmentation network scheme, which is a network that combines dynamic convolutional decomposition feature extraction and multi-scale deformable convolutional techniques (referred to as DYDCNet). This network first introduces dynamic convolutional decomposition based on the attention mechanism and uses a convolutional weight matrix with dynamics to optimize the feature extraction process, which significantly reduces the network parameters and improves the feature extraction efficiency. Subsequently, a deformable convolution technique is used to fuse the null convolution with multiple expansion rates to extend the sensory field and realize feature extraction at different scales. Further, the final segmentation results are refined and optimized by an encoder-decoder architecture. The combination of this series of innovations enables DYDCNet to significantly improve the prediction speed and segmentation accuracy when processing desert region images. Experimental results show that the network has excellent performance on datasets specifically designed for desert features, with an average intersection and merger ratio of 87.75% and an overall accuracy of 91.35%, which outperforms existing mainstream semantic segmentation networks.

INDEX TERMS Desert area, multiscale, dynamic convolution decomposition, deformable convolution, deep learning, ground object classification.

I. INTRODUCTION

The relationship between deserts and oases is interdependent and symbiotic, with oases referring to heterogeneous ecological landscapes that can be maintained in a relatively stable manner, with significant microclimatic effects, based on a large-scale desert background substrate with a small-scale, but sizeable, biological community. Deserts are barren areas where the ground is completely covered by sand, plants are very sparse, rainfall is scarce, and the air is dry. A comprehensive and macroscopic grasp of the spatial distribution pattern of deserts and oases is crucial to the protection of regional ecology, and can achieve a win-win situation for economic development and ecological protection. The oasis is the best part of the arid zone, and because of the complexity of the feature types in the desert area, mapping through manual

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni.

visual interpretation or field surveys requires the consumption of a large amount of resources [1]. With the continuous progress and enrichment of satellite manufacturing technology and payload types, the resolution of the available data resources in the spectral and temporal dimensions has been increasing [2], making it easier to obtain low-cost, fast and high-precision image data in desert areas [1]. However, due to the vast area of the desert region and the complex image characteristics, no current image recognition algorithms are suitable for its geographical characteristics [3], and factors such as atmospheric interference and feature characteristics also impact the recognition of features in desert region images, so feature recognition in the desert region is still extremely challenging [4].

Traditional image recognition methods mainly consist of pixel-based threshold segmentation [5], [6], [7], cluster segmentation [8], decision tree classification [9], region-based segmentation [10], and semantic learning using random forest and conditional random field [11] to construct classifiers. These methods are limited to images with uniform gray-scale distribution and more obvious differences between the gray-scale of the recognition target and the background, and although they are relatively simple to operate, they are not able to segment a large amount of semantic information, which greatly challenges the increasing proliferation of remote sensing data [12]. With the introduction of deep learning technology, research in the computer vision field has greatly progressed, and convolutional neural networks have been gradually applied in the image processing field [13] to achieve the semantic segmentation of images at the pixel level [14]. Semantic segmentation is the process of partitioning a provided image into visually meaningful multiple regions before conducting image analysis and visual understanding [15]. It has a wide range of application areas [16], such as scene analysis [17], automated driving [18], biomedical image research [19], and land cover type analysis based on satellite images [20]. To address the increased complexity of image segmentation scenarios, a series of deep learning-based semantic image segmentation methods [21] have emerged.

In 2015, Long and other scholars proposed the full convolutional neural network (FCN) [22]. Based on FCN, scholars later proposed U-Net [23], which uses the same encoder-decoder structure and achieves the fusion of low-level and high-level image features [24] through multilevel jump connections [25]. The SegNet network proposed by Badrinarayanan et al. utilizes the pooling indices that record the maximal response feature positions of the max-pool layer for upsampling [26], which effectively avoids the consumption caused by upsampling in FCN, and then uses the trainable convolutional layer to make the sparse feature map dense and avoids the additional consumption caused by saving feature maps. In desert image records, features usually exhibit multiscale features. Therefore, effective extraction and integration of this multiscale feature information can significantly enhance the learning ability of these features and yield better image segmentation [28], [29]. PSPNet [29] uses a pyramid pooling module to collect hierarchical information and successfully performs multiscale segmentation analysis of image semantics [30]. Wang et al. [31] combined DeepLabV3+ with CRF to make the remote sensing image boundary clearer. Taoyang et al. [32] embedded the convolutional attention mechanism into DeepLabV3+ network structure to reduce the influence of irrelevant features on recognition accuracy. Zhou et al. and his team proposed a multiscale deep contextual convolutional network called MDCCNet, which is capable of integrating feature maps from different hierarchical networks to achieve semantic segmentation [33]. Wang et al. designed a multiscale deep contextual convolutional network called MDCCNet, which is based on the multiscale feature extraction technique. However, the limited number of samples for certain feature types and the overexposure of certain areas have adversely affected the accuracy of desert segmentation. To segment image features in desert regions quickly and accurately, we must further enhance the fusion capability of multiscale information in images [34]. However, current multiscale feature fusion models often require many computations [35], which leads to low training efficiency.

With the wide application of convolutional neural networks in computer vision [36], large amounts of accurately labeled data have become increasingly necessary [37]. Although semantic segmentation networks based on high-resolution remote sensing image data have been in development for a long time, most of the current research is based on existing open datasets. For example, PASCAL-VOC 2012 [38] is a mainstream dataset in computer vision for recognizing objects from multiple visual object classes in realistic scenes. The Cityscapes dataset mainly focuses on urban street scenes [39]. The ADE20K dataset [40] contains more than 20,000 large-scale scene parsing data with 150 target objects. In addition, scholars have also produced a desert road dataset to monitor the impact of sandstorms on traffic arteries in desert areas [41] and a mangrove forest dataset to reduce environmental damage on the Brazilian coast [42]. Deep learning-based image processing methods are widely used in various aspects such as medicine, bridges, traffic, etc., but there are still fewer applications in segmentation of typical ground objects in desert areas, and few datasets that cover the object classes and typical features of scenes in desert areas. A large amount of observational evidence has shown that desertification has been occurring in most parts of the world over the past few decades [43], and desert areas have shown an expansion trend [45], so the semantic recognition of typical ground objects in desert areas can be achieved by using high-resolution remote sensing images [11] and deep learning techniques [45]. This is of great importance and significance for ecological environmental protection and the promotion of economic development in desert areas.

To respond to the above issues, taking into account the need to identify multiple types of features in the context of large deserts, and the complexity of feature types and

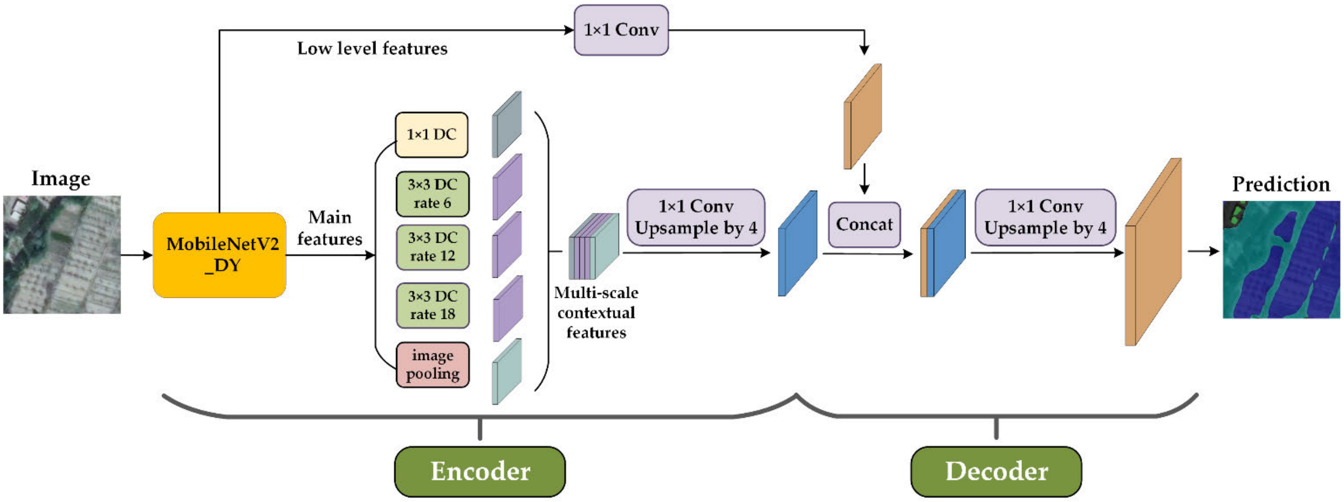

FIGURE 1. The overall structure of DYDCNet.

the irregularity of their location in desert areas, we propose a semantic segmentation model DYDCNet based on the decoding-encoding structure, taking the Xinjiang Ring Tarim Basin region as the study area. The experimental results show that the model exhibits better performance in the semantic segmentation of desert regions datasets than mainstream semantic segmentation networks. Our main contributions are as follows: (1) Using MobileNetV2_DY with a dynamic convolutional decomposition network with dynamic convolutional decomposition as the backbone network can reduce the number of parameters of the model while ensuring feature segmentation accuracy. (2) Aiming to address the characteristics of remote sensing images in desert areas, deformable convolution DC, which integrates multiple expansion rate cavity convolution, is utilized for multiscale contextual information extraction to improve feature extraction in desert areas. (3) A semantic segmentation dataset for the desert area of the Tarim Basin in Xinjiang, China, is produced to provide a database for the recognition and segmentation of typical ground objects in desert areas. The experimental results show that the model exhibits better performance in the semantic segmentation of desert regions datasets than mainstream semantic segmentation networks.

II. MATERIALS AND METHODS

A. DYDCNet ARCHITECTURE

DYDCNet is a deep convolutional neural network [46] with encoding-decoding structure, which optimizes the segmentation result edge accuracy by incorporating an upsampling decoder module, thus greatly improving the accuracy and efficiency of the segmentation results. To accelerate the convergence of the model, we replace the backbone Xception with the lightweight but efficient MobileNetv2 network [47], and to reduce the number of parameters more significantly and obtain higher accuracy, we introduce dynamic convolutional decomposition [48] for the MobileNetv2 network. Concurrently, the DC module can obtain contextual information at different scales to better understand the semantic information of images, the first four parts of the module are deformable convolutions for adaptive learning of receptive fields [49].

Our DYDCNet architecture is shown in Fig. 1. In the encoding stage, the MobileNetV2_DY network with dynamic convolutional decomposition is first used as the backbone network. This network can dynamically adjust the number of layers and channels of the network according to the characteristics of the input image, allowing the network to adapt to different tasks and data distributions with greater flexibility. To achieve adaptive learning of receptive fields for better feature extraction, deformable convolutions are introduced, the DC module, which consists of an averaged pooling layer with globally informative features, a 1 × 1 deformable convolution for raw scale features, and 3 × 3 deformable convolutions with expansions of 6, 12, and 18, respectively. By utilizing three dilated convolutions of different sizes, the module obtains convolution kernels with multiple receptive fields for extracting features at different scales with fewer parameters. Finally, the feature maps extracted by the DC module are concatenated, and the number of channels is compressed by a 1 × 1 convolution. In the decoder stage, the feature map is restored to the original size of the input image through successive upsampling.

B. MobileNetV2_DY MODULE

Dynamic convolution [48] is beneficial for forming lightweight networks and significantly improves performance through its almost negligible computational cost, which motivates its frequent application to vision-related tasks [50]. The core idea of the method is to dynamically integrate multiple convolutional kernels into a single convolutional weight matrix based on the input attention mechanism:

$$W(x) = \sum_{k=1}^{K} \pi_k(x) W_k \ s.t.\ 0 \le \pi_k(x) \le 1, \sum_{k=1}^{K} \pi_k(x) = 1 \tag{1}$$

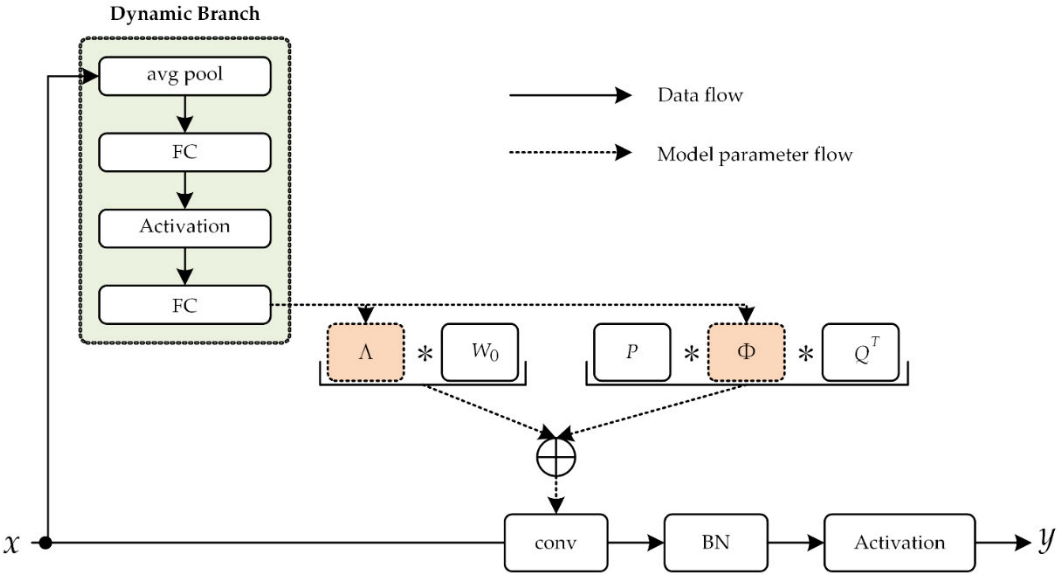

FIGURE 2. Dynamic convolution decomposition layer.

where K is the convolution kernel, $\{W_k\}$ is the attention score, and $\{\pi_k(x)\}$ is the linear aggregation.

However, conventional dynamic convolution faces two main challenges: a lack of compactness due to the use of K kernels and the joint optimization of $\{\pi_k(x)\}$ and the static kernel $\{W_k\}$. These two challenges can be revisited through dynamic convolutional decomposition, while using dynamic channel fusion methods reduces the dimensionality of the latent space and simplifies joint optimization. This makes the network easier to train without sacrificing accuracy.

Fig. 2 shows the dynamic convolutional decomposition layer, where the input x is first dynamically branched to generate $\Lambda(x)$ and $\Phi(x)$, and the convolutional weight matrix $W(x)$ is then generated using (2).

$$W(x) = \Lambda(x) W_0 + P\Phi(x) Q^T \tag{2}$$

where $\Lambda(x)$ is a C×C diagonal matrix. Using this approach, $\Lambda(x)$ implements the channel attention mechanism after the static kernel W_0. The limitations of dynamic convolution are addressed using the dynamic channel fusion mechanism, which is implemented using the full matrix $\Phi(x)$, where each element $\phi_{i,j}(x)$ is a function of x. This approach is employed to drastically reduce the dimensionality of the potential space to construct a more compact model. The dynamic decomposition convolution is achieved with dynamic channel fusion.

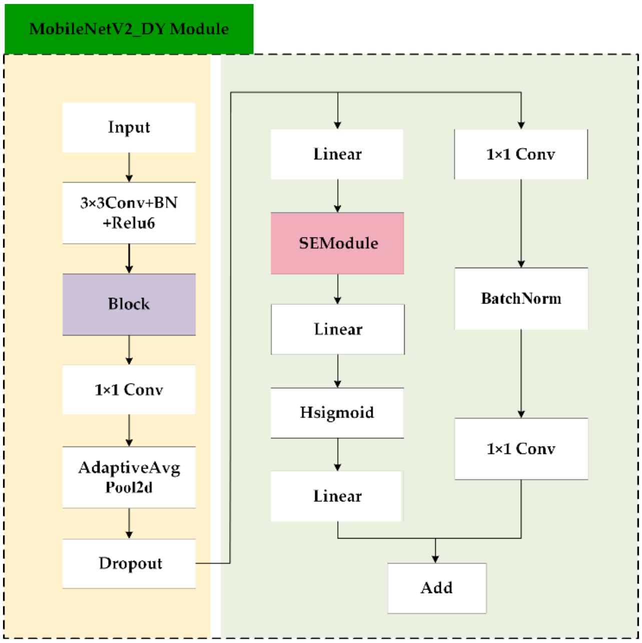

FIGURE 3. The network structure of MobileNetV2_DY.

As shown in Fig. 3, the MobileNetV2_DY we use consists of two main parts: encoding (left) and decoding (right). The encoding part is taken as a 3 × 3 convolution along with a tandem combination of BatchNorm regularization and the ReLU activation function, block module, 1 × 1 convolution, adaptive convolution, and dropout layer. The block module has two cases. If the dynamic convolutional decomposition layer is not used, ordinary convolution is used for feature extraction, and the feature vector is obtained through a series of operations of 3 × 3 convolution with a step size of 1, 1 × 1 convolution operation, BatchNorm regularization, and ReLU6 activation function. If the dynamic convolutional decomposition layer is used, the module begins by using the 1 × 1 convolution operation to adjust the feature dimensions, and the feature vector is obtained after undergoing adaptive pooling, linear linearization, adaptive convolution, BatchNorm regularization, Hsigmoid activation function and SEModule channel fusion module, where the Hsigmoid activation function is a nonlinear function that increases the expressive power of the network. Feature multiplication is

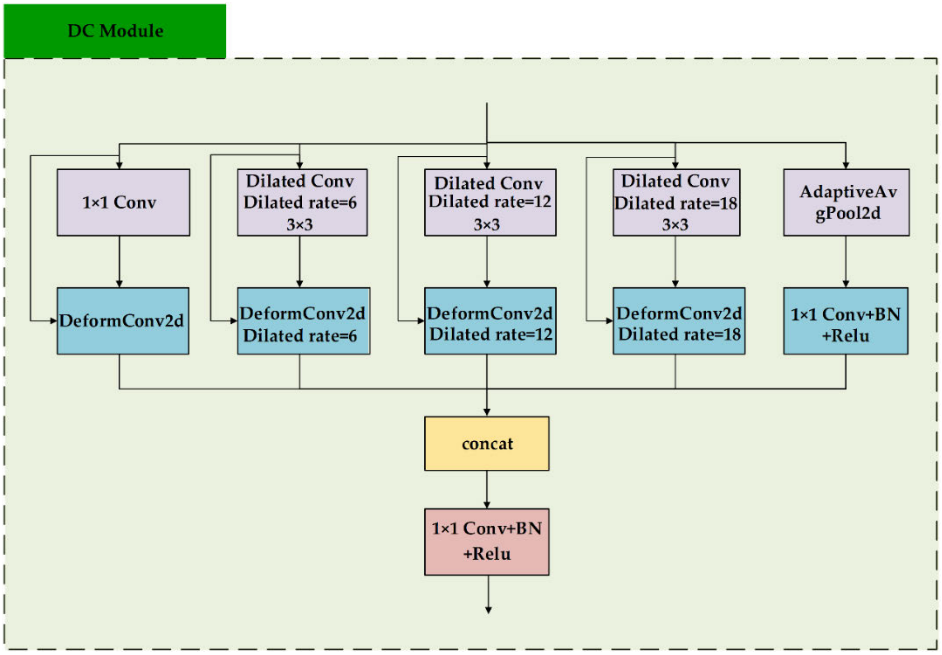
DC Module
1×1 Conv
Dilated Conv Dilated rate=6 3×3
Dilated Conv Dilated rate=12 3×3
Dilated Conv Dilated rate=18 3×3
AdaptiveAvgPool2d
DeformConv2d
DeformConv2d Dilated rate=6
DeformConv2d Dilated rate=12
DeformConv2d Dilated rate=18
1×1 Conv+BN +Relu
concat
1×1 Conv+BN +Relu

FIGURE 4. The network structure of DC.

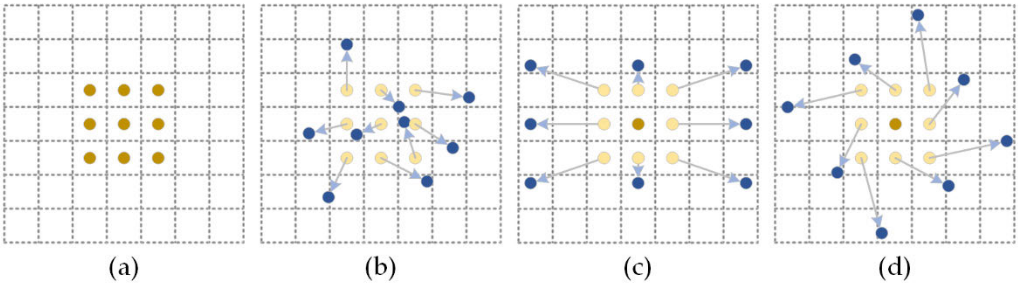

FIGURE 5. Schematic representation of the sampling positions: (a) standard convolution using a regular grid; (b) deformable convolution: deformed sampling positions (dark blue dots) with enhanced offsets (light blue arrows); (c)(d) denote the special case of (b).

used in SEModule to directly transfer the features from the front layer to the back layer. This feature addresses gradient vanishing in the deep network and makes the model easier to train and optimize. Finally, in the decoding part, the features obtained after processing in the encoding part are further optimized and processed.

C. DC MODULE

The DC module has five branches, as shown in Fig. 4. The first branch is composed of a series of 1 × 1 convolution and deformable convolution, and the original feature information and the features after the 1 × 1 convolution operation are used as inputs to the deformable convolution. The second, third and fourth branches are composed of a series of hollow convolution and deformable convolution. At this time, the dilation rates of hollow convolution in the second, third, and fourth branches are 6, 12 and 18, respectively. The hollow convolution can expand the sensory field while guaranteeing the resolution, allowing the feature characteristics of the desert area to be grasped comprehensively and macroscopically. In addition, setting different dilation rates provides the network with different sensory field sizes, i.e., multiscale contextual information is obtained, which benefits the feature recognition of remote sensing images in desert areas with large sizes and irregular distributions of feature locations. The fifth branch adopts adaptive convolution, which consists of a series of 1 × 1 convolution, BatchNorm regularization and the ReLU activation function. Finally, the five branches are concat-enated and spliced in the feature dimension by the concatenation operation. The number of channels is adjusted using 1 × 1 convolution, after which the BatchNorm regularization and ReLU activation function are incorporated to obtain the feature map with high semantic features.

Since the convolution kernel in a standard convolutional neural network has a fixed geometric form, its ability to model geometric transformations is limited. As shown in Fig. 5, the standard convolutional convolution kernel is a fixed rectangular shape, while the deformable convolutional

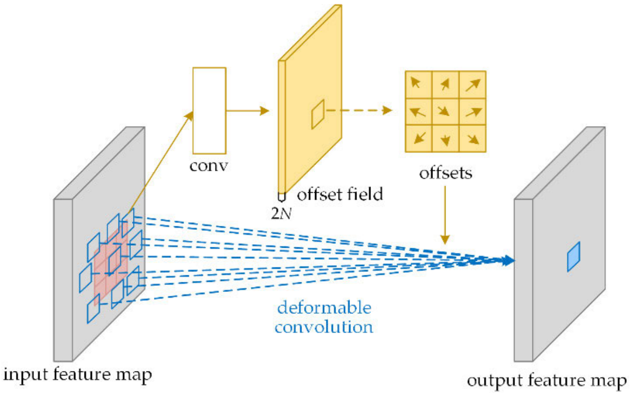

FIGURE 6. **Illustration of 3 × 3 deformable convolution.**

kernel learns an offset at each sampling point to adapt to the geometry of the object. In Fig. 5a, the standard convolutional sampling shape is a fixed rectangle; in Fig. 5b, the sampling position varies according to the offset; in Fig. 5c, scale transformation is achieved; and Fig. 5d depicts a special case for achieving rotation.

While the standard convolution samples pass through a fixed grid R, with each sample point passing through the convolution kernel for weighting, the deformable convolution adds an offset to the sampling:

$$R = \{(-1,-1), (-1,0), \cdots, (0,1), (1,1)\} \tag{3}$$

where R is the regular grid of convolution kernels, where each element represents the offset of all positions of the convolution kernel with respect to the center position, while:

$$\mathrm{y}\left(\mathrm{p}_0\right) = \sum_{\mathrm{p}_\mathrm{n}\in\mathrm{R}} \mathrm{w}\left(\mathrm{p}_\mathrm{n}\right)\cdot\mathrm{x}\left(\mathrm{p}_0+\mathrm{p}_\mathrm{n}\right) \tag{4}$$

where p_0 is the pixel position of the feature map, x is the input feature map, w denotes the weight of the sampled position, and p_n denotes the position in R. $x\ (p_n)$ denotes the pixel value of x at point p_n.

After the input feature map x is sampled, the regular grid R is augmented with offsets $\{\Delta p_n | n = 1, 2, \ldots, N\}$, where N denotes the number of sampling points.

$$\mathrm{y}\left(\mathrm{p}_0\right) = \sum_{\mathrm{p}_\mathrm{n}\in\mathrm{R}} \mathrm{w}\left(\mathrm{p}_\mathrm{n}\right)\cdot\mathrm{x}\left(\mathrm{p}_0+\mathrm{p}_\mathrm{n}+\Delta\mathrm{p}_\mathrm{n}\right) \tag{5}$$

where Δp_n is usually a fraction and represents the offset of the p_n position, so the pixel value of the input feature map x cannot be obtained directly. This value is often obtained using the bilinear difference algorithm with the expression:

$$\mathrm{x(p)} = \sum_{\mathrm{q}} \mathrm{G(q,p)}\cdot\mathrm{x(q)} \tag{6}$$

where $p = p_0 + p_n + \Delta p_n$, denotes an arbitrary position, q denotes a spatial position in the input feature map x, and $G(\cdot,\cdot)$ denotes a bilinear interpolation kernel and is partitioned into two one-dimensional kernels:

$$\mathrm{G(q,p)} = \mathrm{g}\left(\mathrm{q_x},\mathrm{p_x}\right)\cdot\mathrm{g}\left(\mathrm{q_y},\mathrm{p_y}\right) \tag{7}$$

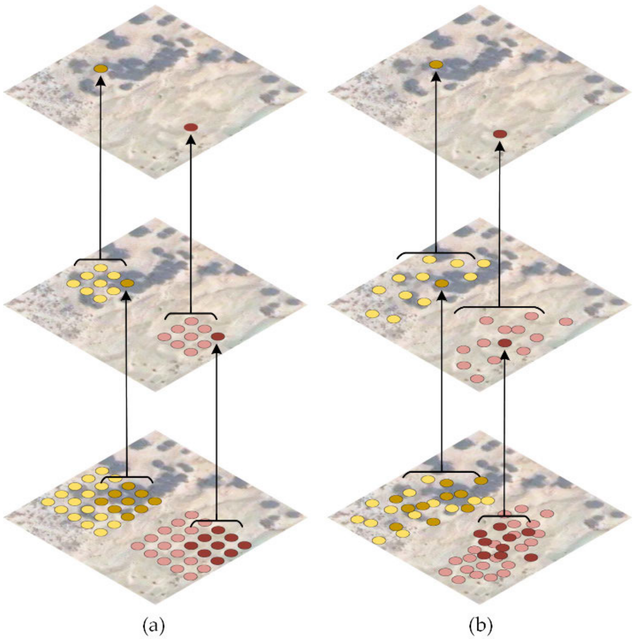

FIGURE 7. **Two representations of two receptive fields in a homebrew dataset: (a) fixed receptive fields in standard convolution, (b) adaptive receptive fields in deformable convolution.**

where $g(a,b) = max\,(0, 1-|a-b|)$.

As shown in Fig. 6, a convolutional layer can be used to determine the offsets. The convolutional kernel is consistent with the existing convolutional layer in terms of spatial resolution and expansion, and each image is assigned different depth information, with numerous channels of dimension 2 N corresponding to N 2D offsets. During training, the convolutional kernel used to generate the output features and the offsets are learned simultaneously. Their gradients are back-propagated to learn the offsets through the bilinear operations in (6) and (7). As shown in Fig. 7a, deformable convolution can adaptively learn the sensory field [51]. In standard convolution, the receptive fields and sampling locations are fixed on the upper feature map, whereas deformable convolution adaptively adjusts to the scale and shape of the target (Fig. 7b). Therefore, for complex targets, deformable convolution has a strong adaptive extraction ability.

D. LOSS FUNCTION

The loss function [52] is a crucial component in the training and validation process of deep learning semantic segmentation models [53], and is the core of backpropagation algorithms [54]. It can be used to quantify the differences between predicted and annotated images, thereby updating parameters [55] to achieve model optimization. In response to the problem of imbalanced samples and high similarity in texture features between buildings and ground in desert datasets, a combination of Focal Loss and Dice Loss was selected to calculate the loss. The definition is as follows:

$$Loss = \lambda Focal\ loss + \left(1-\lambda\right) Dice\ Loss \tag{8}$$

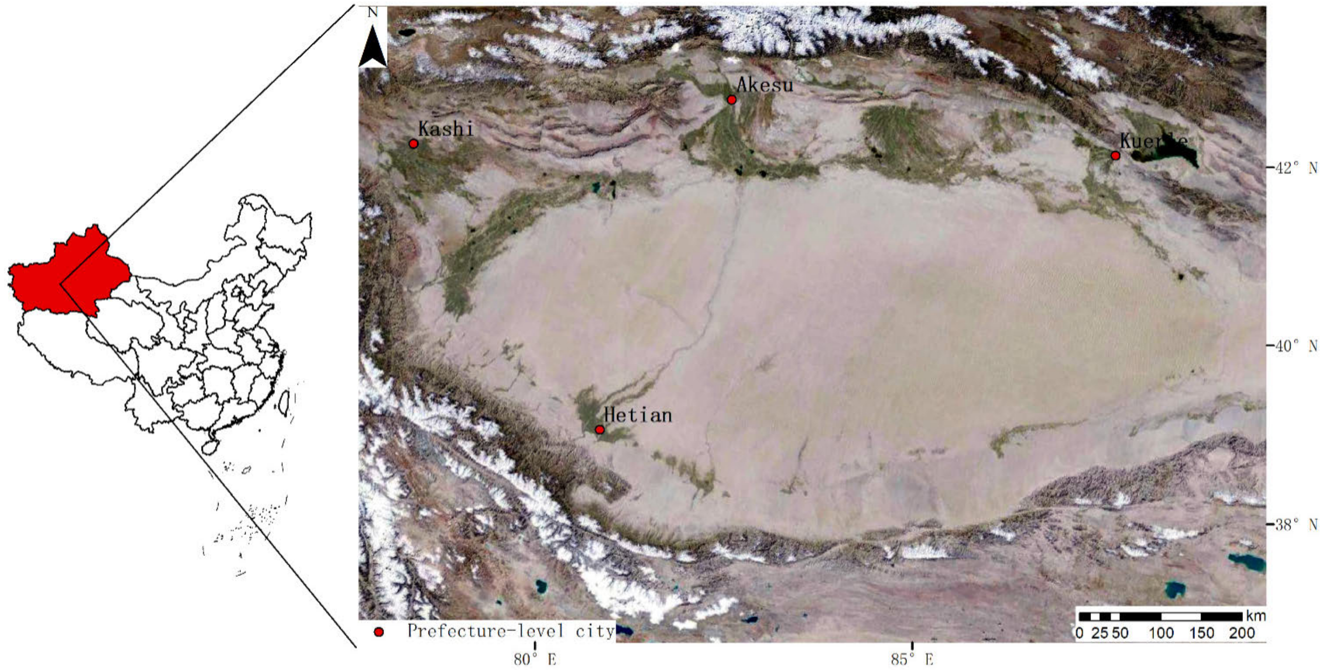

FIGURE 8. **Overview map of the study area.**

1) DICE LOSS

Dice Loss [56] is more sensitive to datasets with imbalanced categories, focusing on the proportion of overlapping parts and the similarity near the boundaries, making it more robust to fuzzy or unclear boundaries. The calculation of Dice Loss is based on the Dice coefficient [57], which is an indicator used to measure the similarity between the model's prediction results and the actual annotated image. The definition is as follows:

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{9}$$

where A and B represent the predicted results and the truth labels. The Dice coefficient ranges from 0 to 1, where 1 represents complete overlap and 0 represents no overlap.

Dice Loss is the conversion of the Dice coefficient into a loss function, the calculation formula is as follows:

$$DiceLoss = 1 - \frac{2 \times |A \cap B|}{|A| + |B|} \tag{10}$$

2) FOCAL LOSS

The Focal Loss can effectively address the issue of category imbalance to a certain extent [58] and prioritize samples that pose challenges in classification, thereby enhancing model performance. The introduction of the balance factor α_t can significantly alleviate this imbalance and enable training to focus more on challenging categories. The calculation formula is as follows:

$$\mathrm{FL}\left(p_t\right) = -\alpha_t \cdot \left(1 - p_t\right)^{\gamma} \cdot \log\left(p_t\right) \tag{11}$$

where p_t represents the predictive probability of the model for a pixel belonging to the correct class, α_t denotes the balancing factor, and γ serves as a regulatory factor used to adjust the weight of challenging-to-classify samples.

III. RESULTS

A. EXPERIMENT DATA

Our main study area is the Tarim Basin region in Xinjiang, China, as shown in Fig. 8, which is far from the ocean and deep inland and is not easily reached by ocean currents, resulting in a distinct temperate continental climate [59]. The Tarim Basin is located in the southern part of Xinjiang and is ring-shaped, with vast deserts in the interior and oases at the edges, so the land type samples in this region are rich. The main sample classes include deserts, the Gobi, oasis farmland, water bodies, and many urban areas; this region is thus more diverse than the regions in other remote sensing land-cover utilization datasets.

The remote sensing image data downloaded from Mapbox in Bigemap. The data we use is a true-color RGB image with three bands acquired in June 2022 from DigitalGlobe's QuickBird, the sensor is a push-scan imaging scanner. True-color images are good for land-cover classification because the colors of the features in such images are very close to or consistent with the colors of the actual features, which makes it easier to manually annotate them and discriminate the category of the features. We used the 0.59 m optical remote sensing image to outline the typical feature area by manual identification, as shown in Fig. 9. The area contains a total of seven ground object types, including desert, Gobi, farmland, buildings, water bodies, bare area and vegetation,

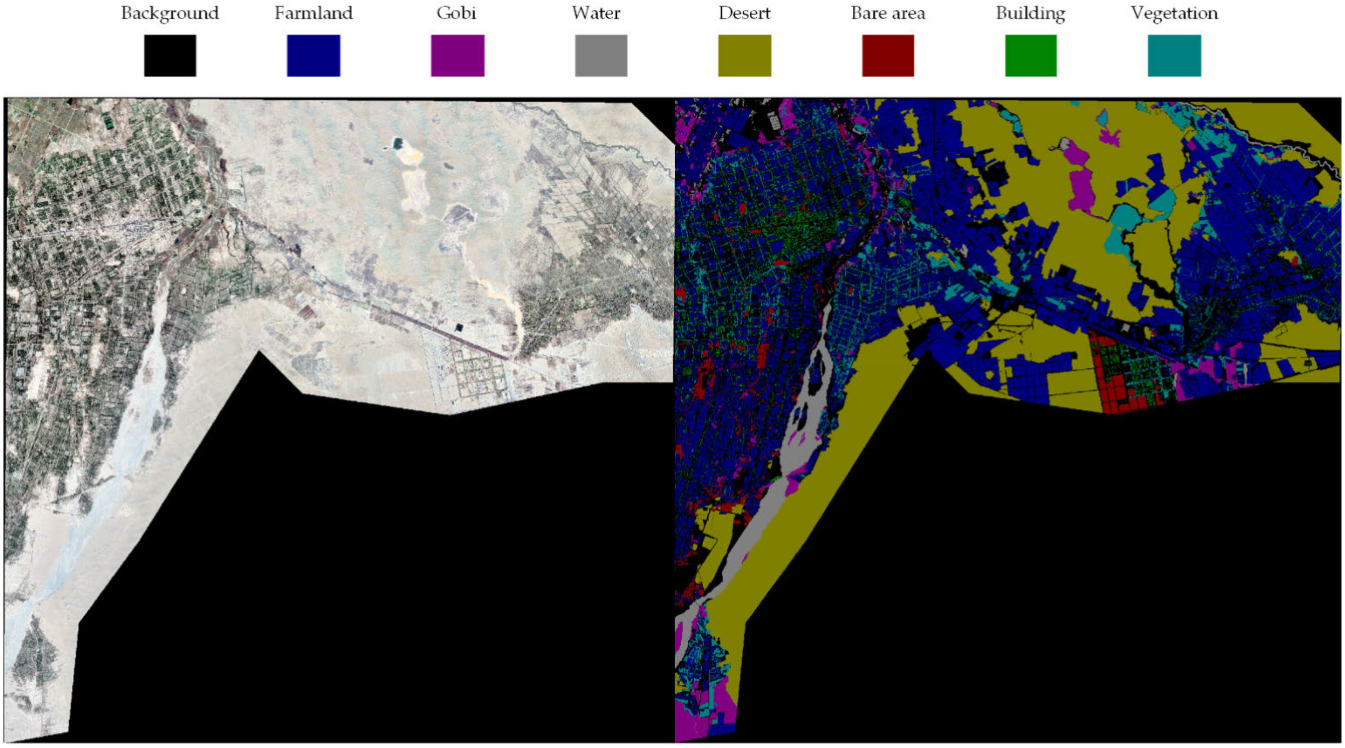

FIGURE 9. Desert dataset in the area around Tarim Basin, Xinjiang.

and the sample area is output as a 256 × 256 image and its corresponding labels, with a total of 20596 image pairs. To enhance the model robustness, the data are enhanced by random flipping and rotation [60]. In addition, the data are divided into a training set, a validation set and a test set in an 8:1:1 ratio for the following experiments, in which the test set is used only for testing and does not contribute to the model training.

B. EVALUATION METRIC

We utilize the mean intersection over union (MIOU), overall accuracy (OA), and G-mean as the evaluation metrics of the model prediction results to quantitatively assess the segmentation performances of the different models. The MIOU represents the mean value after the ratios of intersections and mergers between the prediction results and the true values are summed for each category. The OA represents the sum of correctly classified pixels divided by the total number of pixels. For both of these metrics, a higher value denotes a better model performance [61]. G-mean is an evaluation criterion that combines the values of the two indexes, and when the G-mean value is higher, it indicates a better modal performance.

$$MIOU = \frac{1}{N}\sum_{i=1}^{N}\frac{p_{ii}}{\sum_{j=1}^{N}p_{ij}+\sum_{j=1}^{N}p_{ji}-p_{ii}} \tag{12}$$

TABLE 1. The specific experiment environment.

Type	Environment	Detail
Software	Framework	PyTorch
	Language	Python
	Programming	PyCharm
	Cuda	Cuda 11.1
Hardware	CPU	AMD Ryzen 9 5950X 16-core (3.4 GHz)
	GPU	Nvidia GeForce RTX 3090 (24 GB)
	RAM	128 GB

$$OA = \frac{\sum_{i=1}^{N}p_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N}p_{ij}} \tag{13}$$

$$G-mean = \sqrt{\Pr ecision \times Recall} \tag{14}$$

where N denotes the number of categories to be recognized, p_{ij} denotes the number of class i pixels that are predicted as class j, p_{ii} denotes the number of correctly predicted pixels, and p_{ji} denotes the number of class j pixels that are predicted as class i.

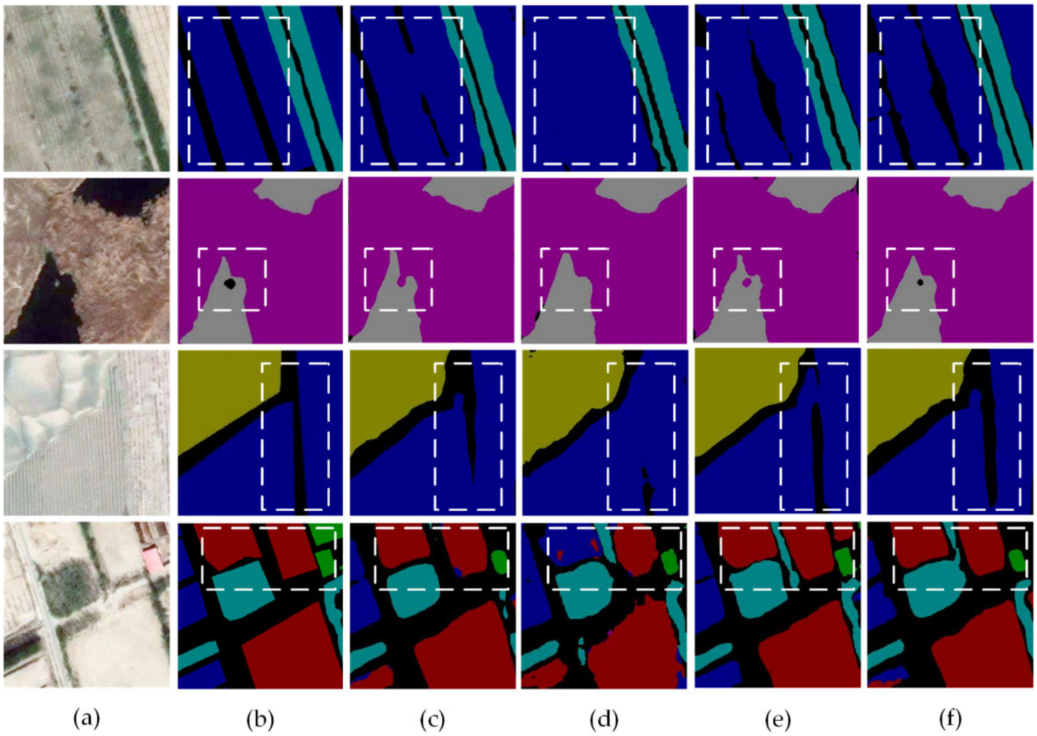

FIGURE 10. Comparison of results of ablation experiments in desert dataset: (a) Original image,(b) Ground truth, (c) Net1,(d) Net2,(e) Net3,(f) DYDCNet.

TABLE 2. Ablation study of DYDCNet.

Methods	baseline	MobileNetV2_DY	Deformable Convolution	MIOU(%)	OA (%)	FPS
Net1	√			84.34	85.3	93
Net2	√	√		84.87	86.56	96
Net3	√		√	85.69	87.39	90
DYDCNet(Ours)	√	√	√	87.75	91.35	106

C. EXPERIMENT ENVIRONMENT

Our created dataset of the desert region of the Ring Tarim Basin is input into the semantic segmentation model for training and validation, with the specific configurations detailed in TABLE 1. For all experiments, the framework is PyTorch, the programming language is Python, and the programming environment is PyCharm. For the hyperparameters, the momentum size is set to 0.9, the batch size to 128, and the decay mode to cosine decay. To facilitate a fair comparison of the experimental results of different models, all the models are trained using the stochastic gradient descent method.

D. EXPERIMENT RESULTS

1) ABLATION EXPERIMENTS

Ablation comparison experiments can be used to verify the usability of each module. To examine the performances of the MobileNetV2_DY module and the DC module, we designed four variations of our network: a network with the MobileNetV2_DY module and DC module removed (Net1), the network with the MobileNetV2_DY module (Net2), the network with the DC module (Net3), and the network with the MobileNetV2_DY module and DC module (DYDCNet). Ablation experiments verified the usability of these modules.

From the segmentation results of the above ablation experiments, we selected four images to analyze. These images are depicted in Fig. 10. For the first and third images, the semantic segmentation results of Net1 and Net3 have clearer classification boundaries than those of Net2. These networks can thus can successfully distinguish different categories in the first and third images, such as farmland and vegetation or desert and farmland. For the second image, only DYDCNet can accurately recognize the nonwater body part inside the water body. Net3 cannot accurately recognize the nonwater body part, although it can separate the nonwater body part from the water body. In this fourth image, Net2 has a large degree of misclassification and omission, and the segmentation result of the boundary of the feature is not sufficiently clear. Net1 and Net3 have better classification effects and refine the boundary part, but they also exhibit misclassification and omission to a certain extent, and DYDCNet achieves

TABLE 3. Segmentation performance of different ModeLS.

Methods	MIOU(%)	OA (%)	FPS	Inference time	G-mean
UNet	73.95	85.69	77	12.98	0.69
RefineNet	70.32	83.40	81	12.34	0.63
PSPNet	85.89	91.2	97	10.3	0.75
DenseASPP	83.6	89.69	85	11.76	0.79
DANet	84.36	89.95	75	13.33	0.76
DYDCNet (Ours)	87.75	91.35	106	9.43	0.81

TABLE 4. Land types segmentation IOU of different models.

Methods	Farmland	Gobi	Water	Desert	Bare area	Building	Vegetation	Background
UNet	72.35	73.68	87.66	85.73	62.15	67.42	70.38	72.23
RefineNet	70.84	67.33	78.59	89.62	63.28	68.51	75.67	68.72
PSPNet	88.73	90.62	95.04	88.27	84.55	74.97	82.75	82.19
DenseASPP	84.61	85.53	93.47	86.76	77.34	73.93	78.64	88.52
DANet	87.35	87.42	92.68	87.86	83.65	74.07	69.72	89.13
DYDCNet (Ours)	89.17	88.45	96.33	94.35	86.77	75.38	83.62	87.93

a better segmentation performance in the multiclassification task of this image. The experimental results show that Net3 clearly outperforms Net2, but the classification effect of using only one of the modules is not satisfactory for some classified images. DYDCNet, however, combines the advantages of the MobileNetV2_DY and DC modules and achieves the best classification results.

TABLE 2 presents a quantitative comparison of the different modules. FPS represents the number of images that the model can process per second [1], which reflects the prediction speed of the model. The DC module achieves a significantly higher accuracy than the MobileNetV2_DY module. Specifically, the MIOU and OA results of the former are 0.82% and 0.83%, respectively, higher than those of the latter, which reflects the DC module's effectiveness in segmenting the features of the desert dataset. Finally, the DYDCNet combining the Mo-bileNetV2_DY module and DC module achieved the best experimental results, showing that each module of DYDCNet is essential for obtaining the best segmentation results.

2) COMPARISON OF EXPERIMENT RESULTS

We chose the following networks for our comparative experiments.

UNet: The feature map obtained from the backbone feature extraction network in the encoding stage is fused with the enhanced feature extraction part in the decoding stage.

RefineNet: The feature map generated in the encoding stage and the output of decoding in the previous stage are simultaneously used as inputs to the RefineNet module, making the fusion of multi-scale features more in-depth.

PSPNet: This is a more extensive network based on spatial pyramid pooling, which is proposed to be able to aggregate global context information of different image regions, and provides a pyramid pooling module to fuse features at different levels.

DenseASPP: This network combines the advantages of parallel and cascaded use of null convolutional layers to produce features of greater scale over a larger area, with increasingly larger receptive field accessible through a series of null convolutions.

DANet: The Dilated ResNet is used as the backbone to feed the resulting feature maps into the two positional attention modules and the channel attention module, and finally to summarise the output features of the two attention modules.

The results are depicted in TABLE 3. The MIOU of our network is 87.75%, which is 13.8%, 17.43%, 1.86%, 4.15% and 3.39% higher than those of the traditional UNet, RefineNet, PSPNet, DenseASPP and DANet models, respectively. The OA of our network is 91.35%, which is 5.66%, 7.95%, 0.15%, 1.66% and 1.4% higher than those of the traditional UNet, RefineNet, PSPNet, DenseASPP and DANet models, G-mean also achieved better results. Furthermore, the DYDCNet more efficiently processes images than the traditional UNet, RefineNet, DenseASPP and DANet. The experimental data show that introducing MobileNetV2_DY with dynamic convolutional decomposition and deformable convolution capable of multiscale feature extraction into our network can enhance the expressive ability of the network and yield better segmentation results. Moreover, utilizing MobileNetV2_DY as the backbone feature extraction module reduces the amount of parameter computations in the model, thus improving the computational speed.

TABLE 4 shows the results of land type segmentation for each network. Relatively high IOU values were achieved

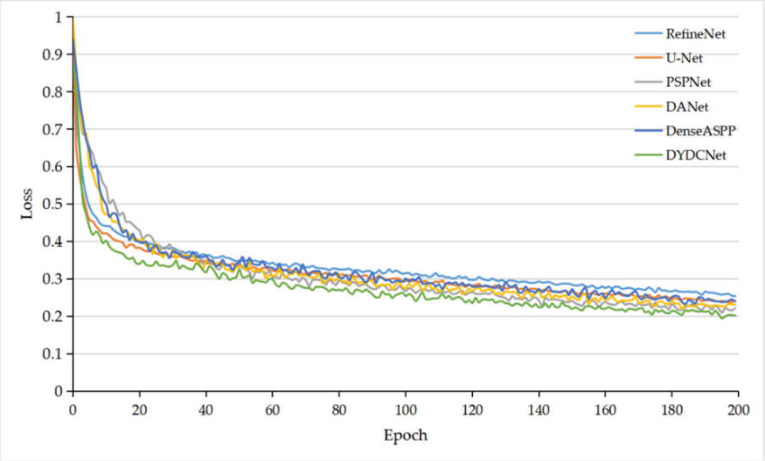

FIGURE 11. Training loss in desert dataset.

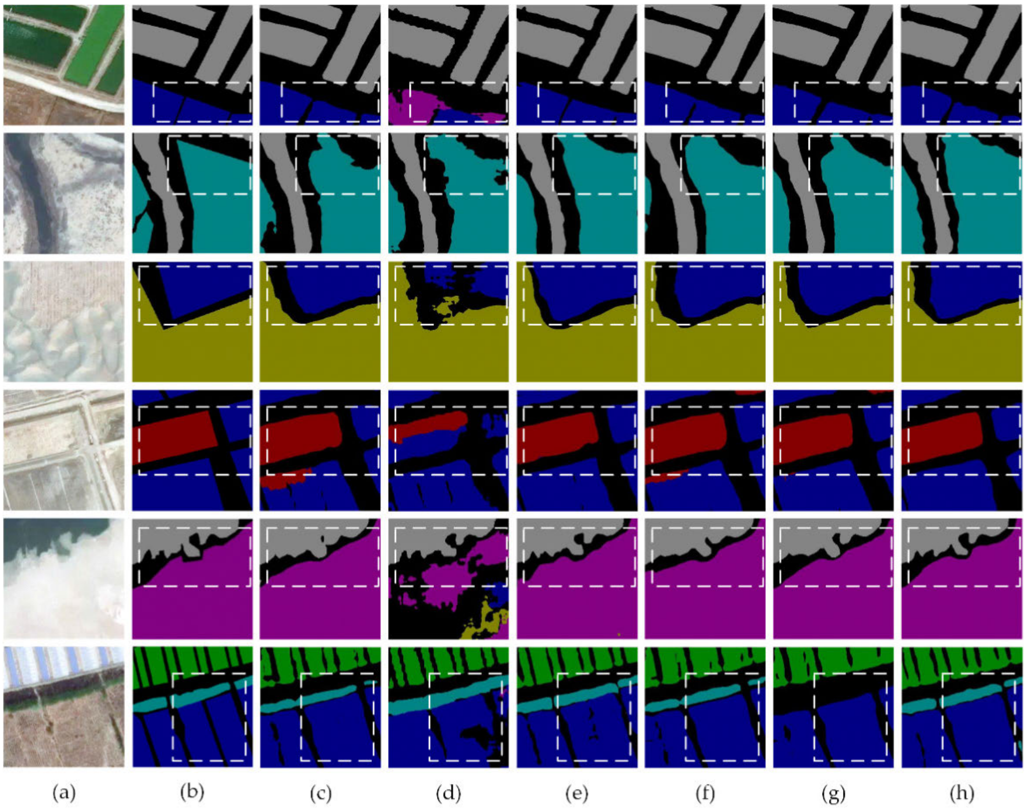

FIGURE 12. Comparison of segmentation results of different models in desert dataset (a) Original image,(b) Ground truth,(c)UNet,(d)RefineNet,(e)PSPNet,(f)Dense ASPP,(g)DANet,(h) DYDCNet.

for all categories except for the buildings category, which suggests that this land type is easier to recognize when there are sufficient samples. DYDCNet's segmentation results for water bodies and deserts exhibited IOU values greater than 90%, at 96.33% and 94.35%, respectively. Producing the dataset revealed that most of the buildings were similar in color to the ground, so the total number of samples for buildings was low, which resulted in a generally lower IOU indicator for this feature type. DYDCNet's IOU for the building category was 7.96% higher than that of UNet, reaching 75.38%. The results show that DYDCNet can better learn the semantic features of the building samples and achieve more accurate results when the number of training samples is small and feature confusion is mitigated. Fig. 11 illustrates the loss during training. After Epoch 170, the loss of the six network models stabilizes, and RefineNet has the highest loss value overall, while DYDCNet has the lowest loss value.

To simply and intuitively analyze the segmentation effects of UNet, RefineNet, PSPNet, DenseASPP, DANet, and our

DYDCNet on the desert dataset from multiple perspectives, we selected six graphs from the prediction results for comparison. Fig. 12 depicts the image, the real labeled map, and the segmentation results of each model for each image. The first map contains a large area of surface water and a small amount of agricultural land, and the fifth map contains a large Gobi region and a small amount of surface water. The maps can be used to validate the prediction results of the models under the category imbalance condition. The second and third maps contain water and Gobi and farmland and desert, respectively. These maps exhibit more distinct feature boundaries and can be used to validate the precision of the model's feature boundary identification. The fourth image contains easily confused boundaries, such as bare area and farmland, which can be used to verify the segmentation performance of the model for images with easily confused features. The sixth image contains multiple categories of buildings, vegetation and farmland, which can be used to compare the classification performances of models in complex scenes.

In the first and fifth figures, water, farmland and the Gobi are better identified. However, RefineNet detected only part of the farmland and the Gobi, and it yielded was a large area of missegmentation, which indicates that extracting features using only dilated convolution is prone to important information loss, which leads to pixel-level misclassification. In the second and third figures, no missegmentation occurs for any of the methods. The segmentation of vegetation by UNet and RefineNet in the second figure and the segmentation of desert and farmland by RefineNet in the third figure are not completely accurate and have unclear boundaries. In contrast, DYDCNet can accurately recognize the feature types and achieves clear feature boundaries. In the fourth figure, bare area and farmland have similar color and texture characteristics, making these categories easier to confuse. The missegmentation yielded by UNet and RefineNet is more obvious. DenseASPP and DANet exhibit slight missegmentation phenomena, identifying a small portion of the edge part of the farmland as bare area. However, PSPNet and DYDCNet exhibit accurate segmentation, so the DYDCNet network can more effectively learn the difference between the two types of features. In the sixth figure the scene is more complex due to the inclusion of multiple types of features. DANet does not accurately recognize the vegetation; RefineNet recognizes the vegetation but cannot completely recognize the farmland; and UNet, PSPNet and Dense ASPP can accurately recognize various types of ground objects. DYDCNet has a better segmentation effect than the other networks, and it can fully perceive the different feature types. DYDCNet combines the MobileNetV2_DY and DC module. MobileNetV2_DY can enhance the expression ability by using the Hsigmoid nonlinear activation function for better feature extraction, and the DC module has a stronger extraction ability. Therefore, DYDCNet is more suitable for the specific environment and feature distribution characteristics of desert areas.

IV. DISCUSSION

Desert remote sensing images are usually characterized by large sizes and irregular feature distributions. The DYDCNet incorporates the MobileNetV2_DY module and DC module and combines the backbone feature extraction network with adaptive feature changes and deformable convolution with different expansion rates. This addresses the issues that feature classes with similar texture and color features are easily confused, exhibit unclear boundaries, and have complex geomorphic features that are difficult to identify. DYDCNet is able to successfully recognize different typical ground object types in the desert dataset and achieves an average intersection and merger ratio (MIOU) and overall accuracy (OA) of 87.75% and 91.35%, respectively, as well as good training efficiency. Fig. 10 and Fig. 12 illustrate that DYDCNet achieved good segmentation results in the desert dataset, especially for small complex images such as buildings and bare area. Compared with previous studies, our approach achieves significant improvements in dealing with feature recognition in desert regions. Previous studies are often limited by the similarity of feature characteristics and complex geomorphological structures, resulting in poor recognition accuracy or blurred boundaries. In summary, the DYDCNet model provides an effective solution in addressing the challenges in remote sensing image analysis in desert areas.

However, there are some limitations to our work. In our experiments, we found that although the proposed method has yielded significant progress in remote sensing image segmentation in desert areas, it can still be further improved and optimized. Therefore, in future research, we can focus on data diversity, including more remote sensing data under geographic locations, seasons, and light conditions. This is conducive to the better adaptation of the model to image segmentation in desert areas under different contexts and improves the generalizability of the model, which will help to better reveal the spatial patterns and ecosystems of desert areas and provide support for environmental protection and sustainable development.

V. CONCLUSION

Highly accurate segmentation results of remote sensing images of desert areas with clear boundaries can help us understand the current situation and ecological environment of desert areas in a timely manner. Based on the encoding-decoding structure and multiscale feature extraction, we propose DYDCNet, a network that fuses MobileNetV2_DY and multiscale deformable convolutional DC, to further improve the effect of semantic segmentation in desert areas. First, we used the circum-Tarim Basin area in Xinjiang, China, as the study area. We used manual recognition to outline the feature regions, to produce a semantic segmentation dataset for typical ground objects in the desert region. Then, we conducted experiments and a comparative analysis of the results based on this dataset. The experimental results show that DYDCNet achieves clearer classification boundaries by refining the boundary part of the features using the encoding-decoding structure while ensuring segmentation efficiency. Multiscale feature extraction by deformable convolution results in better segmentation performance in multi-classification tasks. When the number of building category samples is small and its features are easily confused with other features, DYDCNet can better learn the specific features of building samples and obtain more accurate feature segmentation results. In summary, our improved network is an effective automatic segmentation method for typical ground objects in desert region images.

REFERENCES

[1] L. Wang, L. Weng, M. Xia, J. Liu, and H. Lin, "Multi-resolution supervision network with an adaptive weighted loss for desert segmentation," *Remote Sens.*, vol. 13, no. 11, p. 2054, May 2021, doi: 10.3390/rs13112054.

[2] B. Zhang, Y. Wu, B. Zhao, J. Chanussot, D. Hong, J. Yao, and L. Gao, "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1814–1822, 2022, doi: 10.1109/JSTARS.2022.3148139.

[3] M. Xia, Y. Cui, Y. Zhang, Y. Xu, J. Liu, and Y. Xu, "DAU-Net: A novel water areas segmentation structure for remote sensing image," *Int. J. Remote Sens.*, vol. 42, no. 7, pp. 2594–2621, Apr. 2021, doi: 10.1080/01431161.2020.1856964.

[4] L. Weng, L. Wang, M. Xia, H. Shen, J. Liu, and Y. Xu, "Desert classification based on a multi-scale residual network with an attention mechanism," *Geosci. J.*, vol. 25, no. 3, pp. 387–399, Jun. 2021, doi: 10.1007/s12303-020-0022-y.

[5] H. Y. Wang, D. L. Pan, and D. S. Xia, "A fast algorithm for two-dimensional Otsu adaptive threshold algorithm," *Acta Autom. Sinica*, vol. 33, no. 9, pp. 969–970, 2005, doi: 10.1360/aas-007-0968.

[6] K. Bhargavi and S. Jyothi, "A survey on threshold based segmentation technique in image processing," *Int. J. Innov. Res. Develop.*, vol. 3, no. 12, pp. 234–239, 2014.

[7] L. Qingge, R. Zheng, X. Zhao, W. Song, and P. Yang, "An improved Otsu threshold segmentation algorithm," *Int. J. Comput. Sci. Eng.*, vol. 22, no. 1, pp. 146–153, 2020, doi: 10.1504/ijcse.2020.10029225.

[8] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade*, 2nd ed. Berlin, Germany: Springer, 2012, pp. 561–580.

[9] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[10] I. Mehidi, D. E. C. Belkhiat, and D. Jabri, "An improved clustering method based on k-means algorithm for MRI brain tumor segmentation," in *Proc. 6th Int. Conf. Image Signal Process. their Appl. (ISPA)*, Nov. 2019, pp. 1–6, doi: 10.1109/ISPA48434.2019.8966891.

[11] X. Sun, X. Lin, S. Shen, and Z. Hu, "High-resolution remote sensing data classification over urban areas using random forest ensemble and fully connected conditional random field," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 8, p. 245, Aug. 2017, doi: 10.3390/ijgi6080245.

[12] J. Fan, M. Zhang, J. Chen, J. Zuo, Z. Shi, and M. Ji, "Building change detection with deep learning by fusing spectral and texture features of multisource remote sensing images: A GF-1 and sentinel 2B data case," *Remote Sens.*, vol. 15, no. 9, p. 2351, Apr. 2023, doi: 10.3390/rs15092351.

[13] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–11, Dec. 2019, doi: 10.1186/s13640-019-0417-8.

[14] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1089–1106, Aug. 2019, doi: 10.1007/s10462-018-9641-3.

[15] X.-F. Wang, D.-S. Huang, and H. Xu, "An efficient local Chan–Vese model for image segmentation," *Pattern Recognit.*, vol. 43, no. 3, pp. 603–618, Mar. 2010, doi: 10.1016/j.patcog.2009.08.002.

[16] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*.

[17] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015, doi: 10.1109/TCSVT.2014.2358029.

[18] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 954–960.

[19] I. Rizwan I Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," *Informat. Med. Unlocked*, vol. 18, May 2020, Art. no. 100297, doi: 10.1016/j.imu.2020.100297.

[20] S. I. Toure, D. A. Stow, H.-C. Shih, J. Weeks, and D. Lopez-Carr, "Land cover and land use change analysis using multi-spatial resolution data and object-based image analysis," *Remote Sens. Environ.*, vol. 210, pp. 259–268, Jun. 2018, doi: 10.1016/j.rse.2018.03.023.

[21] X. Yu, J. Fan, J. Chen, P. Zhang, Y. Zhou, and L. Han, "NestNet: A multiscale convolutional neural network for remote sensing image change detection," *Int. J. Remote Sens.*, vol. 42, no. 13, pp. 4898–4921, Jul. 2021, doi: 10.1080/01431161.2021.1906982.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440, doi: 10.1109/CVPR.2015.7298965.

[23] 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. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[24] D. Yang, G. Liu, M. Ren, B. Xu, and J. Wang, "A multi-scale feature fusion method based on U-Net for retinal vessel segmentation," *Entropy*, vol. 22, no. 8, p. 811, Jul. 2020, doi: 10.3390/e22080811.

[25] X. Meng, P. Wang, H. Yan, L. Xu, J. Guo, and Y. Fan, "Multi-graph convolution network with jump connection for event detection," in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 744–751.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.

[27] D. Duarte, F. Nex, N. Kerle, and G. Vosselman, "Multi-resolution feature fusion for image classification of building damages with convolutional neural networks," *Remote Sens.*, vol. 10, no. 10, p. 1636, Oct. 2018, doi: 10.3390/rs10101636.

[28] X. Song, S. Jiang, and L. Herranz, "Multi-scale multi-feature context modeling for scene recognition in the semantic manifold," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2721–2735, Jun. 2017, doi: 10.1109/TIP.2017.2686017.

[29] A. Krishna, "A distributed scheme for accelerating and scaling PSPNet," M.S. thesis, Dept. Comput. Sci., Kidder Hall, 2021.

[30] Y. Kong, Y. Liu, B. Yan, H. Leung, and X. Peng, "A novel DeepLabV3+ network for SAR imagery semantic segmentation based on the potential energy loss function of Gibbs distribution," *Remote Sens.*, vol. 13, no. 3, p. 454, Jan. 2021, doi: 10.3390/rs13030454.

[31] J. Q. Wang, J. S. Li, H. C. Zhou, and X. Zhang, "A typical element extraction method for remote sensing images based on DeepLabV3+ with CRF," *Comput. Eng.*, vol. 45, no. 10, pp. 260–265, 2019, doi: 10.19678/j.issn.1000-3428.0053359.

[32] M. Taoyang, Z. Wei, H. Xiaoyu, and L. Dan, "A Rice fall recognition method based on improved DeepLabV3+ model combined with UAV remote sensing," *J. China Agricult. Univ.*, vol. 27, pp. 143–154, Jul. 2022.

[33] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, and L. J. Latecki, "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, no. 2, pp. 555–570, Mar. 2019, doi: 10.1007/s11280-018-0556-3.

[34] I. Heidarpour Shahrezaei and H.-C. Kim, "Fractal analysis and texture classification of high-frequency multiplicative noise in SAR sea-ice images based on a transform-domain image decomposition method," *IEEE Access*, vol. 8, pp. 40198–40223, 2020, doi: 10.1109/ACCESS.2020.2976815.

[35] M. Xia, W. Liu, K. Wang, W. Song, C. Chen, and Y. Li, "Non-intrusive load disaggregation based on composite deep long short-term memory network," *Expert Syst. Appl.*, vol. 160, Dec. 2020, Art. no. 113669, doi: 10.1016/j.eswa.2020.113669.

[36] H. Li, X. Fan, L. Jiao, W. Cao, X. Zhou, and L. Wang, "A high performance FPGA-based accelerator for large-scale convolutional neural networks," in *Proc. 26th Int. Conf. Field Program. Log. Appl. (FPL)*, Aug. 2016, pp. 1–9, doi: 10.1109/FPL.2016.7577308.

[37] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, ''The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243, doi: 10.1109/CVPR.2016.352.

[38] M. Everingham and J. Winn, ''The PASCAL visual object classes challenge 2012 (VOC2012) development kit,'' *Pattern Anal., Stat. Model. Comput. Learn.*, vol. 2007, nos. 1–45, p. 5, 2012, doi: 10.1007/11736790_8.

[39] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, ''The cityscapes dataset,'' in *Proc. CVPR Workshop Future Datasets Vis.*, Jun. 2015, pp. 1–2.

[40] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, ''Semantic understanding of scenes through the ADE20K dataset,'' *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019, doi: 10.1007/s11263-018-1140-0.

[41] X. Zhang, C. Li, S. Shao, and Y. Peng, ''The extraction of the desert roads based on the U-Net network from remote sensing images,'' *Proc. SPIE*, vol. 12129, pp. 87–94, Dec. 2021, doi: 10.1117/12.2625577.

[42] G. M. D. S. Moreno, O. A. D. C. Júnior, O. L. F. de Carvalho, and T. C. Andrade, ''Deep semantic segmentation of mangroves in Brazil combining spatial, temporal, and polarization data from Sentinel-1 time series,'' *Ocean Coastal Manage.*, vol. 231, Jan. 2023, Art. no. 106381, doi: 10.1016/j.ocecoaman.2022.106381.

[43] P. A. Dirmeyer and J. Shukla, ''The effect on regional and global climate of expansion of the world's deserts,'' *Quart. J. Roy. Meteorolog. Soc.*, vol. 122, no. 530, pp. 451–482, Jan. 1996, doi: 10.1002/qj.49712253008.

[44] Y. Chen, H. Lu, H. Wu, J. Wang, and N. Lyu, ''Global desert variation under climatic impact during 1982–2020,'' *Sci. China Earth Sci.*, vol. 66, no. 5, pp. 1062–1071, May 2023, doi: 10.1007/s11430-022-1052-1.

[45] M. Coccia, ''Deep learning technology for improving cancer care in society: New directions in cancer imaging driven by artificial intelligence,'' *Technol. Soc.*, vol. 60, Feb. 2020, Art. no. 101198, doi: 10.1016/j.techsoc.2019.101198.

[46] 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. (ECCV)*, 2018, pp. 801–818.

[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, ''MobileNetV2: Inverted residuals and linear bottlenecks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[48] Y. Li, Y. Chen, X. Dai, M. Liu, D. Chen, Y. Yu, L. Yuan, Z. Liu, M. Chen, and N. Vasconcelos, ''Revisiting dynamic convolution via matrix decomposition,'' 2021, *arXiv:2103.08756*.

[49] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, ''Deformable convolutional networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[50] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, *arXiv:1704.04861*.

[51] F. Chen, F. Wu, J. Xu, G. Gao, Q. Ge, and X.-Y. Jing, ''Adaptive deformable convolutional network,'' *Neurocomputing*, vol. 453, pp. 853–864, Sep. 2021, doi: 10.1016/j.neucom.2020.06.128.

[52] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, ''A comprehensive survey of loss functions in machine learning,'' *Ann. Data Sci.*, vol. 9, no. 2, pp. 187–212, Apr. 2022.

[53] S. Hao, Y. Zhou, and Y. Guo, ''A brief survey on semantic segmentation with deep learning,'' *Neurocomputing*, vol. 406, pp. 302–321, Sep. 2020.

[54] L. Kirsch and J. Schmidhuber, ''Meta learning backpropagation and improving it,'' in *Proc. Neural Inf. Process. Syst.*, 2021, pp. 1–10.

[55] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, ''A survey of quantization methods for efficient neural network inference,'' 2021, *arXiv:2103.13630*.

[56] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, ''Dice loss for data-imbalanced NLP tasks,'' 2019, *arXiv:1911.02855*.

[57] B. Prencipe, N. Altini, G. D. Cascarano, A. Brunetti, A. Guerriero, and V. Bevilacqua, ''Focal dice loss-based V-Net for liver segments classification,'' *Appl. Sci.*, vol. 12, no. 7, p. 3247, Mar. 2022.

[58] K. Pasupa, S. Vatathanavaro, and S. Tungjitnob, ''Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification,'' *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 11, pp. 15259–15275, Nov. 2023.

[59] H. Sun, Y. Chen, W. Li, F. Li, Y. Chen, X. Hao, and Y. Yang, ''Variation and abrupt change of climate in Ili River Basin, Xinjiang,'' *J. Geograph. Sci.*, vol. 20, no. 5, pp. 652–666, Oct. 2010.

[60] J. Chen, J. Fan, M. Zhang, Y. Zhou, and C. Shen, ''MSF-Net: A multiscale supervised fusion network for building change detection in high-resolution remote sensing images,'' *IEEE Access*, vol. 10, pp. 30925–30938, 2022, doi: 10.1109/ACCESS.2022.3160163.

[61] Z. Wang, J. Wang, K. Yang, L. Wang, F. Su, and X. Chen, ''Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with DeepLabV3+,'' *Comput. Geosci.*, vol. 158, Jan. 2022, Art. no. 104969, doi: 10.1016/j.cageo.2021.104969.

JUNFU FAN is currently a Professor with the Department of Surveying and Mapping Engineering, Shandong University of Technology. His research interests include high-performance geocomputing, spatial analysis algorithms, GIS, and remote sensing of urban environments.

YU GAO was born in Zaozhuang, Shandong, China, in 1997. She received the B.S. degree in surveying and mapping engineering from Shandong University of Technology, Zibo, China, in 2019, where she is currently pursuing the master's degree. Her research interests include intelligent geographic computing and semantic segmentation algorithms based on high-resolution remote-sensing images.

ZONGWEN SHI was born in Weifang, Shandong, China, in 1999. He received the B.S. degree in surveying and mapping engineering from Shandong University of Technology, Zibo, China, in 2021, where he is currently pursuing the master's degree. His research interests include intelligent geographic computing and semantic segmentation algorithms based on high-resolution remote sensing images, spatial analysis algorithm design, and development.

PING LI was born in Dezhou, Shandong, China, in 1999. She received the B.S. degree in surveying and mapping engineering from Shandong University of Technology, Zibo, China, in 2022, where she is currently pursuing the master's degree. Her research interests include spatial analysis algorithm design and GeoAI.

GUANGWEI SUN was born in Weifang, Shandong, China, in 1979. He is currently a Professor with the School of Architectural Engineering and Geomatics, Shandong University of Technology. His research interests include spatial analysis algorithms and GIS applications.