Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Globe230k: A Benchmark Dense-Pixel Annotation Dataset for Global Land Cover Mapping

Qian Shi[1,2], Da He[1,2,*], Zhengyu Liu[1,2], Xiaoping Liu[1,2], and Jingqian Xue[1,2]

[1]School of Geography and Planning,Sun Yat-sen University, Guangzhou 510275, China. [2]Guangdong Provincial Key Laboratory for Urbanization and Geo-simulation, Guangzhou 510275, China.

*Address correspondence to: heda@mail.sysu.edu.cn

Citation: Shi Q, He D, Liu Z, Liu X, Xue J. Globe230k: A Benchmark Dense-Pixel Annotation Dataset for Global Land Cover Mapping. *J. Remote Sens.* 2023;3:Article 0078. https://doi.org/10.34133/remotesensing.0078

Submitted 24 February 2023
Accepted 23 August 2023
Published 16 October 2023

Global land cover map provides fundamental information for understanding the relationship between global environmental change and human settlement. With the development of data-driven deep learning theory, semantic segmentation network has largely facilitated the global land cover mapping activity. However, the performance of semantic segmentation network is closely related to the number and quality of training data, and the existing annotation data are usually insufficient in quantity, quality, and spatial resolution, and are usually sampled at local region and lack diversity and variability, making data-driven model difficult to extend to global scale. Therefore, we proposed a large-scale annotation dataset (Globe230k) for semantic segmentation of remote sensing image, which has 3 superiorities: (a) large scale: the Globe230k dataset includes 232,819 annotated images with a size of 512 × 512 and a spatial resolution of 1 m, including 10 first-level categories; (b) rich diversity: the annotated images are sampled from worldwide regions, with coverage area of over 60,000 km^2, indicating a high variability and diversity; (c) multimodal: the Globe230k dataset not only contains RGB bands but also includes other important features for Earth system research, such as normalized differential vegetation index (NDVI), digital elevation model (DEM), vertical–vertical polarization (VV) bands, and vertical–horizontal polarization (VH) bands, which can facilitate the multimodal data fusion research. We used the Globe230k dataset to test several state-of-the-art semantic segmentation algorithms and found that it is able to evaluate algorithms in multiple aspects that are crucial for characterizing land covers, including multiscale modeling, detail reconstruction, and generalization ability. The dataset has been made public and can be used as a benchmark to promote further development of global land cover mapping and semantic segmentation algorithm development.

Introduction

Land use/land cover (LULC) is the link between human activities and the natural environment [1]. Over the past hundred years, the global LULC has undergone unprecedented dramatic changes with the advancement of industrialization and urbanization, causing a series of resource, environmental, and sustainable development problems, such as deforestation [2], flood inundation [3], and climate warming [4]. Therefore, high-frequency and high-resolution monitoring of LULC is urgently necessitated for us to understand their evolution pathway and to make informed decisions to alleviate the impact of human activities on climate and ecological environment, for realizing the sustainable development goals [5,6].

Satellite images have become one of the important data sources for LULC change monitoring due to their long-term monitoring capabilities and high spatial resolution. The past decades have witnessed an era of vigorous progress of global LULC activities based on satellite remote sensing images and automatic classification algorithms, which unveiled a global spatial configuration of land covers. Early LULC products are mainly based on moderate solution imaging spectroradiometer (MODIS) data, ENVISAT-MERIS data, and other coarse-resolution data, e.g., Global Land-cover Classification (GLC2000) product at 1 km [7], Land-cover product (MOD12Q1) at 500 m [8], the European Space Agency (ESA) Climate Change Initiative Land-cover product (CCI_LC) [9], and Global Land-cover map (GlobCover) at 300 m [10]. Recent advances in Landsat, ASTER, SPOT, Sentinel-2 satellite and processing capabilities facilitate-observation to higher resolution, 30-m resolution images quickly predominated in LULC monitoring task [2,11–15].

LULC monitoring is mainly achieved by per-pixel classification of remote sensing image. Traditional classification method is based on the expert interpretation of the electromagnetic signal feature reflected from the land covers, such as the red edge spectral feature and costal spectral feature. However, this method largely depends on domain knowledge, and it is time-consuming and difficult to apply to large-scale area.

Machine learning algorithm is then developed in remote sensing imagery classification community to automatically learn the interpretation rule based on regression learning mechanism, which substantial improves the automation level of classification process [16–21]. It commonly learns a hyperplane in feature space that can distinguish different land covers

based on the handcrafted spectral, texture, or phenology features through supervision, and the trained model can then be used to automatically predict large-scale land cover map. Classical machine learning methods include maximum likelihood estimation [22], logistic regression [23], support vector machines (SVM) [24], decision trees [25], random forests (RF) [26], Markov random field (MRF) [27], and backpropagation neural network (BP) [28] algorithms. For land cover mapping, FROM_GLC [12] was produced by employing 4 classifiers including SVM and RF on Landsat TM, ETM+ data, and time series MODIS enhanced vegetation index (MODIS EVI) dataset; GlobeLand30 [11] was produced by an integrated approach of pixel- and object-based methods; GLC_FCS30 [14] was produced by building a local adaptive RF model. However, the handcrafted feature relies on expert feature engineering, which has limitations and may not be representative. Besides, the handcrafted features usually remain at low level like texture, spectral, etc., which has information gap from the semantic understanding of LULC.

Deep learning theory has made great breakthroughs in the remote sensing community in recent years due to its strong feature representation learning ability [29,30]. Benefiting from its end-to-end architecture, the representative features of land covers can be automatically learned from training sample pairs of remote sensing images and its corresponding ground truth annotations. Deep learning method uses multiple convolutional layers to extract the intrinsic features from the remote sensing images and estimate the LULC label of each pixel, and the ground truth annotations are used to supervise the estimated result and backpropagate the gradient error to update the network parameter, during which process the representative feature can be learned. Compared with the handcrafted features in machine learning, the high-level features are more representative, discriminative, and robust, which make it possible to accommodate to various terrestrial situations.

Semantic segmentation network is mainly responsible for remote sensing image classification task in deep learning community. It is a process of understanding an image at the pixel level, i.e., each pixel in the image is assigned to a class label [31]. Different from recognizing the commercial scene or residential scene in an image, the semantic segmentation network can delineate the boundaries of each land object in the scene and achieve densely pixel-wise predictions both semantically and locationally. This type of network is built based on "encoder–decoder" structure [32]. The encoder is responsible for aggregating context to extract high-level semantic features; the decoder is responsible for upsampling the features to the target size to restore the spatial location; and a classifier is followed to predict the class probability of each pixel. Aiming at how to learn as many as semantic information while preserving the spatial location information, semantic segmentation network has made many improvements, which are systematically reviewed in the "Review of Semantic Segmentation in Earth Observation Community" section. Semantic segmentation has now been widely deployed to remote sensing communities, such as environmental monitoring [16,33], precision agriculture [34,35], tree species mapping [36], and building footprint extraction [37,38]. Besides, multimodal data fusion is also an alternative hot topic of semantic segmentation in remote sensing community, which can provide auxiliary information for better identification of land covers. For example, SAR (synthetic aperture radar) image [vertical–vertical (VV)/vertical–horizontal (VH)] is free from cloud contamination and dense continuous observations for better classification [39], NDVI dataset can help for the classification of forest cover [40], and DEM dataset can benefit the mountainous landscape classification [41].

Training sample is the basis of data-driven deep learning method. It has been proven that generalization performance of network usually relies on the amount and quality of training samples [42]. On consideration of the importance of densely pixel-wise annotation dataset, some researchers have dedicated to increase the amount of the annotations to ensure the generalization of the network model, which we make a comprehensive summarization in Table 1. For example, Shao et al. [43] published a WHDLD dataset with 4,940 images by annotating on 2-m resolution Google Earth images at Wuhan, China, which mainly includes urban scene and covers circa 1,295 km^2. Wang et al. [44] developed the LoveDA dataset with 5,987 images by annotating on 0.3-m Google Earth image at urban and rural area of Wuhan, Nanjing, and Changzhou of China, with coverage of 565 km^2. Li et al. [45] published a WHU-OHS annotation dataset with coverage of 204,341 km^2, which is annotated on 7,795 Zhuhai-1 OHS images over China at 10-m resolution. Tong et al. [46] proposed a GID annotation dataset with coverage of circa 500,000 km^2 by annotating on 4-m resolution Gaofen-2 images.

However, on the one hand, current annotation datasets are limited to some specific regions or countries, and cannot ensure the diversity and variability over the whole world. When applying to unseen regions, semantic segmentation network model that is overfitted on the limited training data may undergo a dramatic performance drop. Although some countries have released annotation datasets, like the Vaihingen dataset and Postdam dataset [47] in German, most of the annotations are not available, such as used in producing NLCD map in American, GlobeLand30 map, and FROMGLC map for global world. On the other hand, the category system is not consistent among different annotation dataset, which cannot be jointly used to train a single network model. Furthermore, many existing datasets are unbalanced in category proportions, i.e., some are only collected in urban areas with fewer samples of bare land, glaciers, and wetlands, and some are only collected in mountain areas without impervious surface samples, making it difficult for network model to transfer to large regions.

In this study, we create a large-scale annotated dataset (Globe230k) for LULC mapping, which is annotated on Google Earth image of 1-m spatial resolution, with more than 3×10^{10} annotated pixels, and covers more than 60,000 km^2 all over the world. The Globe230k dataset is annotated by numerous experts and students major in surveying and mapping after the necessary training, through visual interpretation on very high-resolution images, as well as in situ field survey, under the guidance of the organized annotation pipeline. Three characteristics of the proposed dataset are the following: (a) It is the largest annotated dataset for LULC mapping, which has 232,819 images with a size of 512 × 512 pixels at 1-m resolution, and 10 classes are defined according to the standard category system; (b) the sampling locations are collected across the world to ensure the diversity and comprehensiveness of the annotations. Besides, in order to ensure the category balance, we intentionally give more chance to the rare categories to be sampled, such as wetland and glacier; (c) multimodal data can provide multidimensional information in land cover mapping and has been proved to play an important role in improving the classification accuracy of land cover objects.

Table 1. Comparison of the existing LULC annotation dataset in remote sensing community.

Dataset	Number of patches	Patch size	Category	Resolution	Band
WHDLD [43]	4,940	256 ×256	6	2 m	3
DLRSD [43]	2,100	256×256	17		3
MSLCC [91]	2	5,957×8,149, 6,031 ×5,596	4	10 m	-
Drone Deploy [92]	55	6,000 ×6,000	7	0.1 m	3
GF2 Dataset for 3DFGC [93]	11	1,417×2,652, 1,163×2,120	5	4 m	4
Semantic Drone Dataset [94]	400	6,000 ×4,000	22	-	3
MiniFrance-DFC22 [95]	2,322	2,000×2,000	15	-	-
Postdam [47]	38	6,000 ×6,000	6	0.05 m	4
Vaihingen [47]	33	2,000×2,000	6	0.09 m	4
DeepGlobe Land Cover Classification Challenge [96]	803	2,448 ×2,448	7	0.5 m	3
DLR-SkyScapes [97]	16	5,616 ×3,744	31	-	3
GID Large scale Classification 5 classes [46]	150	7,200 ×6,800	5	4 m	3
GID Fine Land cover Classification 15 classes [46]	10	7,200 ×6,800	15	4 m	3
LoveDA [44]	5,987	1,024×1,024	7	0.3 m	3
WHU-OHS [45]	7,795	512 × 512	24	10 m	32
Globe230k	**232,819**	512 × 512	10	1 m	7

Therefore, instead of the RGB images, we also collected elevation data, vegetation index data, and polarized data to provide multimodal information for promoting data fusion research. Finally, the proposed Globe230k dataset is taken to test several state-of-the-art semantic segmentation networks, to evaluate the effectiveness of this dataset in inspecting the algorithm performance.

The remainder of this paper is organized as follows: The "Annotation Workflow and Dataset" section 2 introduces the specific annotation pipeline and the foundation information of the Globe230k dataset. The background of the semantic segmentation network for retrieving LULC is provided in the "Review of Semantic Segmentation in Earth Observation Community" section. Validation experiment of the state-of-the-art semantic segmentation network models is tested on Globe230k dataset in the "Benchmark Result" section. Discussion and conclusion are provided in the "Conclusion" section.

Annotation Workflow and Dataset

Annotation region selection

The generalization and representativeness of the annotation samples is crucial to reliable semantic segmentation network training. Therefore, we adopt a strategy to determine the sampling locations to ensure that they are as uniform as possible across the global land surface, which can take diverse features into consideration.

We first uniformly sampled point of interest (POI) on the WSG84 ellipsoid and excluded sampling points in nonterrestrial area. Each sampling point generates a sampling region [area of interest (AOI)] of the same size (0.033281°×0.033281°) on the ellipsoid coordinate as the candidate regions for annotation. Then, we downloaded the corresponding 17-level Google Earth high-resolution image (circa 1-m resolution) in each candidate area and then manually screened the images that are identifiable for the subsequent annotation. Specifically, we organized a special screening team to select the images that are clear, without mosaic effect or cloud contamination, and with identifiable categories, as the candidate annotation region.

However, the uniform sampling strategy cannot ensure the balance of each land cover categories, since the real-world land cover category ratio is extremely uneven. From the existing global land cover products (e.g., WorldCover10 and FROMGLC10), it is obvious that most of the land covers within the sampling regions are concentrated in common categories such as forests, grasslands, and cropland, while some categories, such as ice/snow and wetlands, are very rare. The unbalanced ratio of the land cover categories is very detrimental to classification task.

To alleviate this problem, we stratified the sampling regions according to the land cover categories provided by the existing global land cover products and tried to compensate for the rare categories that are almost missing in the preliminary sampling regions. Specifically, we conducted a second-round region sampling to search for the regions that have rare categories indicated by the global land cover products, which was included into the preliminary regions to rebalance the category ratio. Finally, we selected 6,500 candidate regions as well as their corresponding Google Earth images for the subsequent annotation procedure, and their distributions are shown in Fig. 1.

Annotation principal

We perform a comprehensive and reliable manual annotation on the candidate images based on 2 groups of people, i.e.,

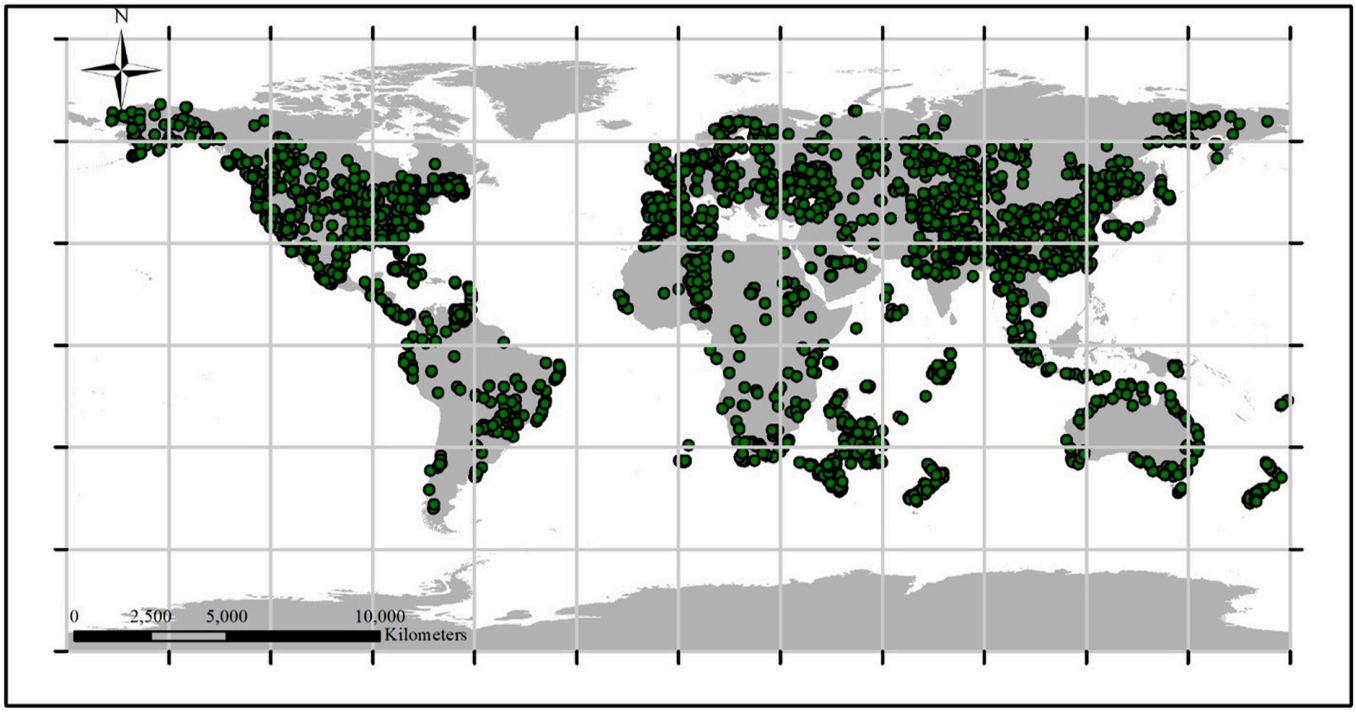

Fig. 1. The global spatial distribution of the sampling POIs.

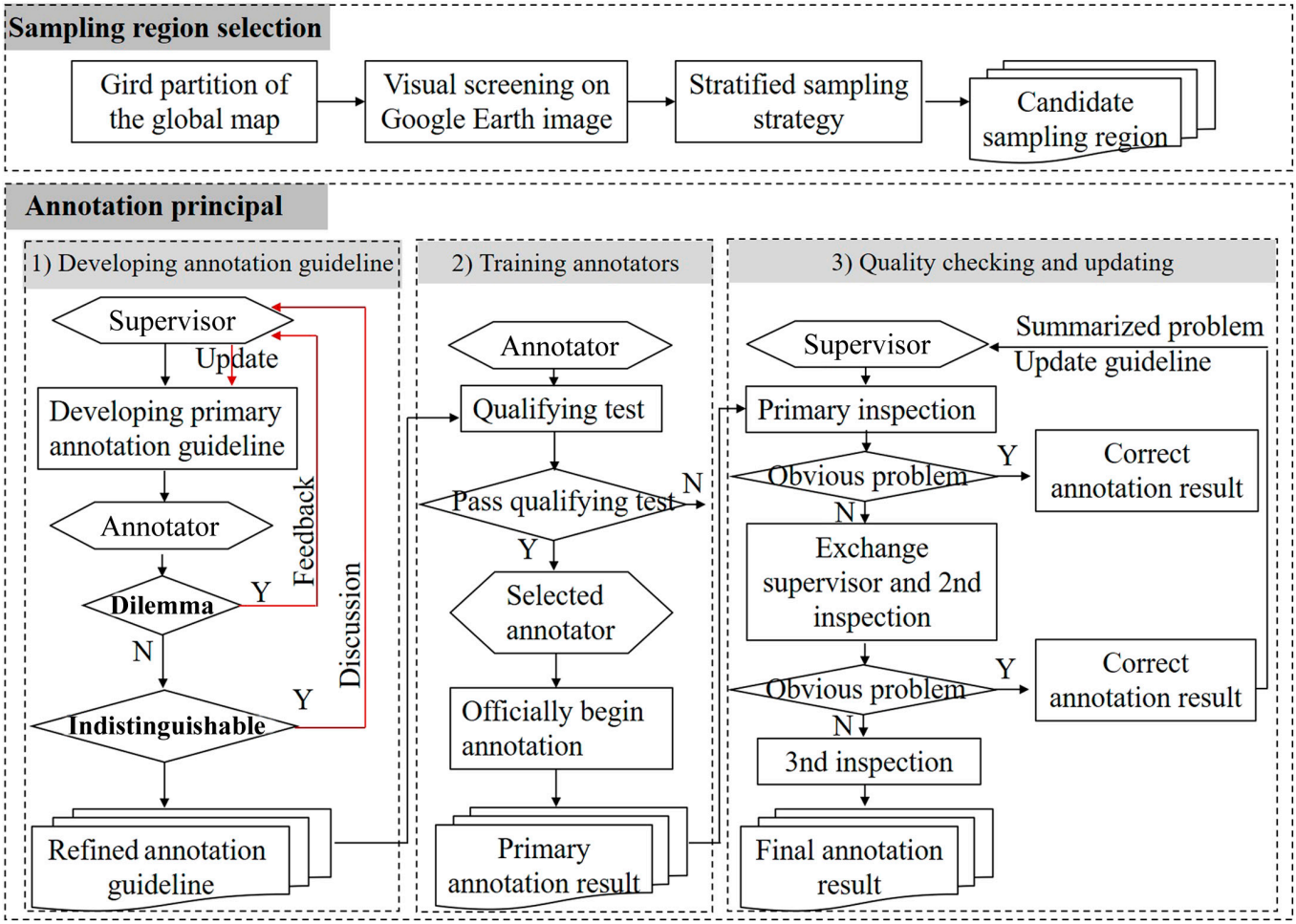

Fig. 2. Pipeline of the annotation principal.

annotators and supervisors, under the standardized annotation guidelines and the strict checking rules. The whole pipeline of the annotation is summarized in Fig. 2, which consists of 3 steps:

(i) Developing an annotation guideline: According to the LULC standard (GB/T21010-2017) and expert opinions, we formulate a preliminary annotation guideline, including the category system, class definition, class instances, and practical principle. In determining the category system, we follow the widely used global land cover product to define 10 classes, i.e., cropland, forest, grass, shrub, wetland, water, tundra, impervious surfaces, bareland, and ice/snow, and the class definitions and class instances can be found in GB/T21010-2017. Six practical principles are formulated:

(a) All clearly identifiable objects in the above 10 categories must be annotated by annotators and cannot be omitted;

(b) The segmentation mask of each land cover object should precisely adhere to its boundary;

(c) When annotating, the image should be zoom-in or zoom-out according to the size of the object, so as to obtain the finest boundaries;

(d) When encountering with the dilemma that the guideline is not applicable to some special situations, the supervisors will provide solutions to the annotators' feedback and update the guideline, to avoid further erroneous labeling results.

(e) Objects identified as indistinguishable or too fragmented will not be annotated after discussion and approval by the supervisors.

(f) All annotation work should be implemented using Octopusboard 3.0, which is a self-developed annotation software.

(ii) Training the annotators: Before an annotator participates in the annotation work, he/she will be professionally trained according to the annotation guidelines. After training, the annotator needs to complete a specified test to examine his/her annotation ability. Only if he/she passes the specified test can we allow the annotator to participate in the annotation work, so as to ensure that the annotation of all images is consistent, complete, and accurate. Noteworthy is that although the very high spatial resolution of the image brings benefits of the clear contours, it also increases the difficulty of outlining the object. Therefore, it averagely takes about 3 h for an annotator to complete one image.

(iii) Multiple rounds of quality checking and updating: After the first round of annotation work is completed, 3 rounds of quality inspection are deployed to ensure the quality of the annotation result:

(a) The supervisors conduct a preliminary inspection; each supervisor is responsible to a fixed group of annotators. For each annotated image, the supervisors will point out the annotation problem for further correction;

(b) After the first round of inspection, each supervisor exchanges the group of annotators and conducts the second round of inspection to correct the remaining problems;

(c) After the second round of inspection and correction, the supervisors should summarize the annotation problems and further update the annotation guideline. After updating, a further round of inspection is implemented, and we obtain the final annotation result.

During each round of inspection, all the annotated objects should be inspected, and problems such as mislabeling, missing labels, and inaccurate boundaries are corrected.

Annotation result

In addition to the Google Earth images and the annotations, auxiliary data at the 6,500 sampling regions are also collected, including the vegetation index, polarization map (SAR), and the digital elevation model (DEM) data, which are aimed to provide data support for multimodal remote sensing research in land cover mapping community.

For auxiliary features, we computed the required data on Google Earth Engine (GEE) and downloaded them for each sampling region. For vegetation index, we adopted the most widely used normalized differential vegetation index (NDVI) calculated by the spectral information in Sentinel-2 image, and used the cloud mask provided by the Sentinel-2 image to exclude the cloud contamination. For polarization map, we directly used the VV and VH bands of the Sentinel-1 image. For DEM data, we downloaded the NASA30mDEM product data. All auxiliary data were sampled to the same resolution with the optical image and kept as close in time as possible.

After the annotation work, we obtained 6,198 land cover annotation images with 1-m spatial resolution, which covers about 60,000 km^2. Through statistics, it is found that the Globe230k dataset has a total of circa 3×10^{10} annotated pixels, and the proportion of the annotated pixels for each category is

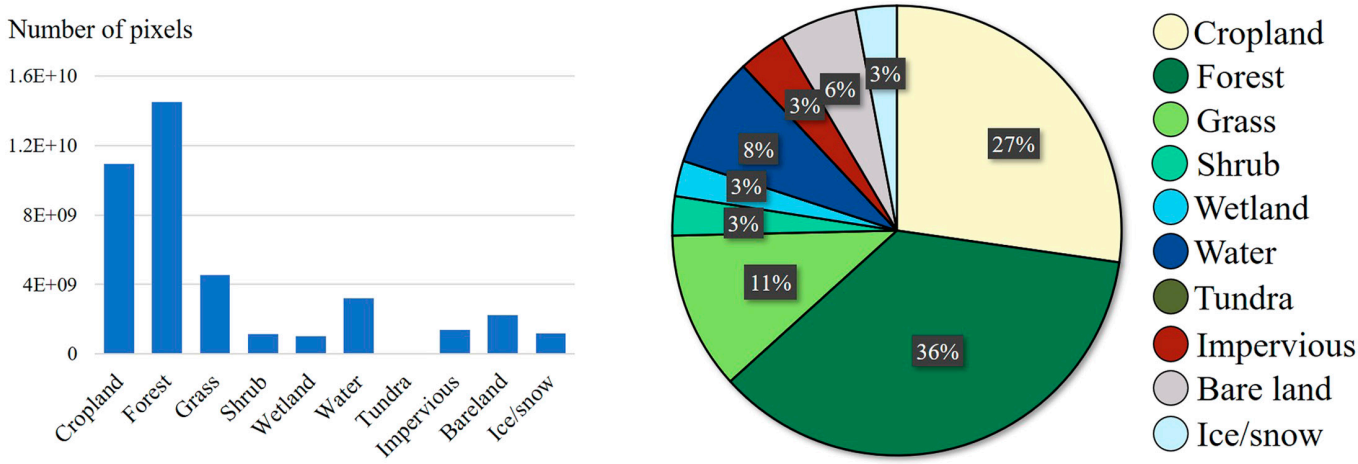

Fig. 3. Statistics of the proportion of annotated pixels for each category.

shown in Fig. 3. It is obvious that benefiting from the rebalance strategy in annotation region selection, the proportion of shrub, ice/snow, and wetland have increased significantly to reach a relative balance, except for tundra, which are difficult to distinguish in optical images. Besides, we make a statistics of spectral reflectance of RGB bands to inspect the interclass separability among the categories (Fig. 4), and find that the 10 classes have considerable difference in spectral feature.

We crop the nonoverlapping patches from the 6,198 annotation images with size of 512 × 512 pixels, resulting in a total of 232,819 patches (some of the patches are displayed in Figs. 5 to 9 for each 10 class). We randomly divide the patches into training set, validation set, and test set with a ratio of 7:1:2.

Review of Semantic Segmentation in Earth Observation Community

With the growing demand of LULC dynamic monitoring, an increasing number of semantic segmentation methods and training datasets for remote sensing image classification have emerged. Semantic segmentation is a process of understanding an image at the pixel level, i.e., each pixel X in the image is assigned to a class label Y [31], which can be demonstrated in Fig. 10. Apart from recognizing the urban scene or rural scene in an image, the boundaries of each land cover object should also be precisely delineated. Therefore, different from image classification task that predicts one label for an image, we need dense pixel-wise predictions both semantically and locationally from semantic segmentation models [48].

This sort of high-level semantic understanding demand cannot be achieved without the context information of each pixel, since geographical objects are closely connected to the surrounding scenes, which can provide cues for the prediction of this pixel. For example, planes are located inside airports, ships are located in harbors, and mangroves generally exist alongside shores. Figure 11 demonstrates a comprehensive example of how context information helps for the semantic recognition. With limited receptive field, the center object is difficult to be identified, while with the global receptive field, the surrounding context such as harbor, water related to the center object, indicates that the center object is highly probably a ship.

However, due to its local connection property, the convolution operation cannot have a global perception of the geographical scene, resulting in a limited receptive field and difficulty in capturing the contextual information that is helpful for the semantic reasoning of each pixel.

In order to extract global spatial context, current semantic segmentation networks usually use stacked multilayer convolution kernels to deepen the network structure, and use pooling operation to aggregate the context information, which can enlarge the receptive field for capturing global context [30,49].

However, the aggregation process of the context inevitably discards the detailed spatial information of each pixel, which contradicts the purpose of spatial location prediction in semantic segmentation, leading to a blurry classification result with vague spatial locations. Therefore, the trend of the semantic segmentation research mainly focuses on how to preserve or

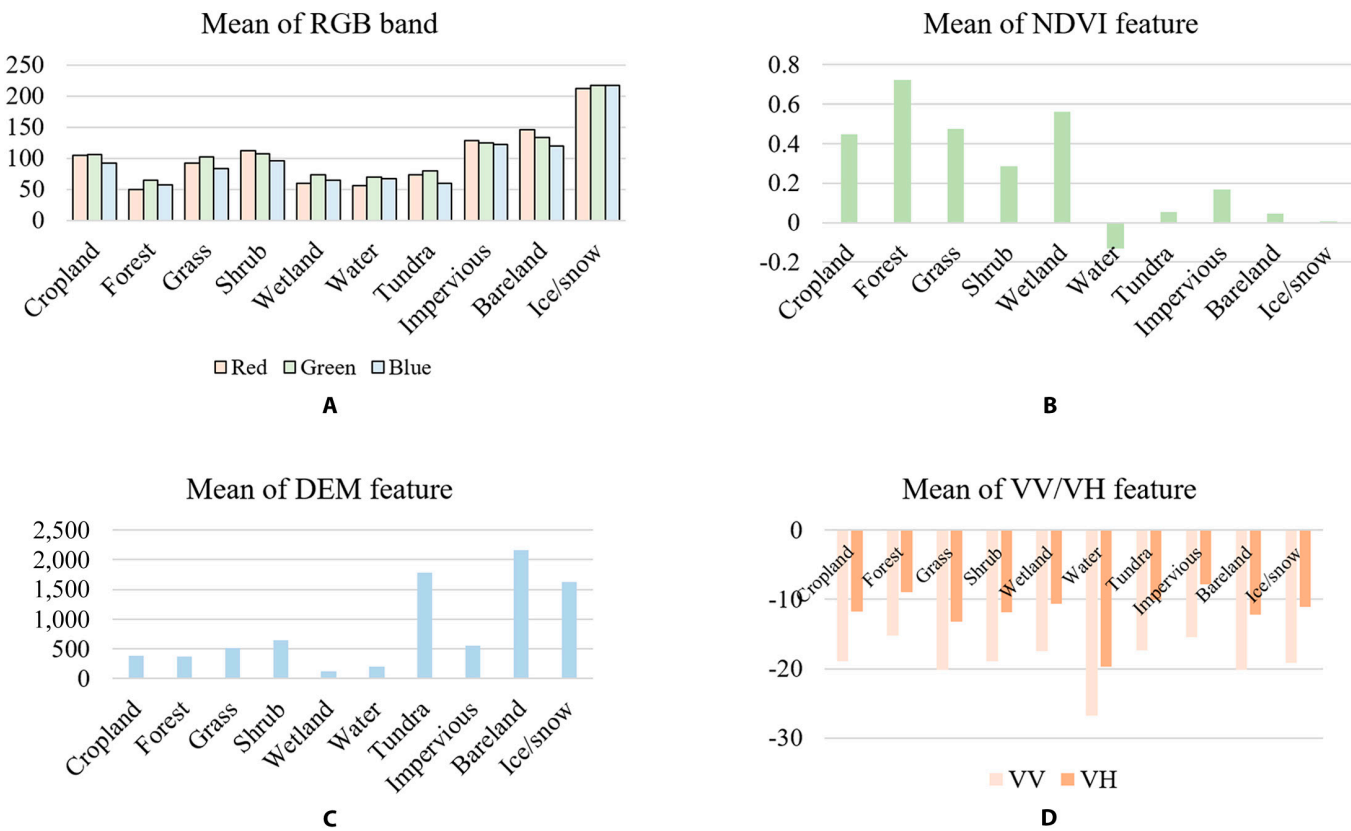

Fig. 4. (A) Statistics of the spectral reflectance of RGB bands for 10 classes. (B) Statistics of the NDVI for 10 classes. (C) Statistics of the DEM for 10 classes. (D) Statistics of the VV/VH for 10 classes.

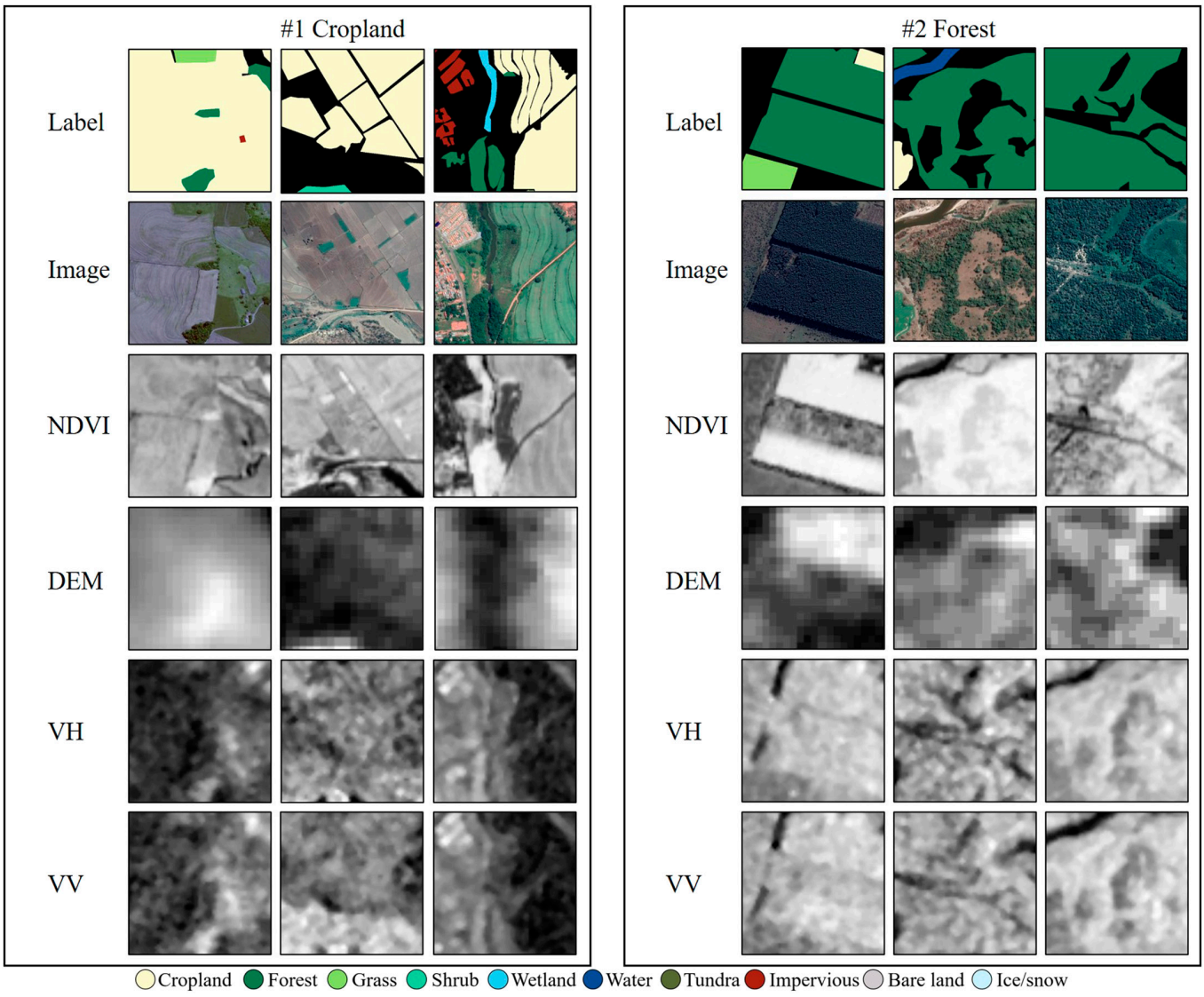

Fig. 5. Examples of the annotated patches of size 512 × 512 for cropland and forest.

restore the detailed location information during the aggregation of the context, or use other methods to increase the receptive field and help better extract high-level semantic information [50,51], which can be systematically categorized into 5 types (Fig. 12): (a) pooling for context extraction and deconvolution for spatial location restoration; (b) dilated convolution for context extraction; (c) multiscale pyramid for context extraction; (d) attention for context extraction; and (e) full self-attention operation. We will introduce these methods in the following sections.

Pooling for context extraction and deconvolution for spatial location restoration

Traditional deep learning approaches for LULC classification adopted patch classification manner, in which each pixel was separately predicted by feeding a patch image centered at this pixel. The main reason of only predicting the center pixel of the input patch is that the classification networks usually use full connected layers at the end of the network, which convert the patch to a single probability vector, and thus ignore the spatial location information [30].

This dilemma motivates the development of the fully convolutional network (FCN) [32]. FCN maintains the spatial information by replacing fully connected layers with deconvolutional layers, which can upsample the coarse feature map to the same size as the input image and restore the spatial location. Due to the advantage of both capturing context and restoring location of each pixel, FCN popularizes the use of end-to-end convolution networks for semantic segmentation task.

However, it is found that the lost location information during the aggregation process of the pooling layer is extremely difficult to be precisely recovered by simple deconvolution. To tackle this problem, Long et al. [32] discovered that the low-level features such as texture or location usually exist in the shallow layer of the encoder part, and thus designed a skip connection strategy to directly transmit the low-level features to the decoder part to inject location cues into the deconvolution layers.

FCN laid the "encoder–decoder" structure foundation for the semantic segmentation network research, i.e., the encoder part is aimed to extract contextual information by shrinking the feature map size, and the decoder part is aimed to restore

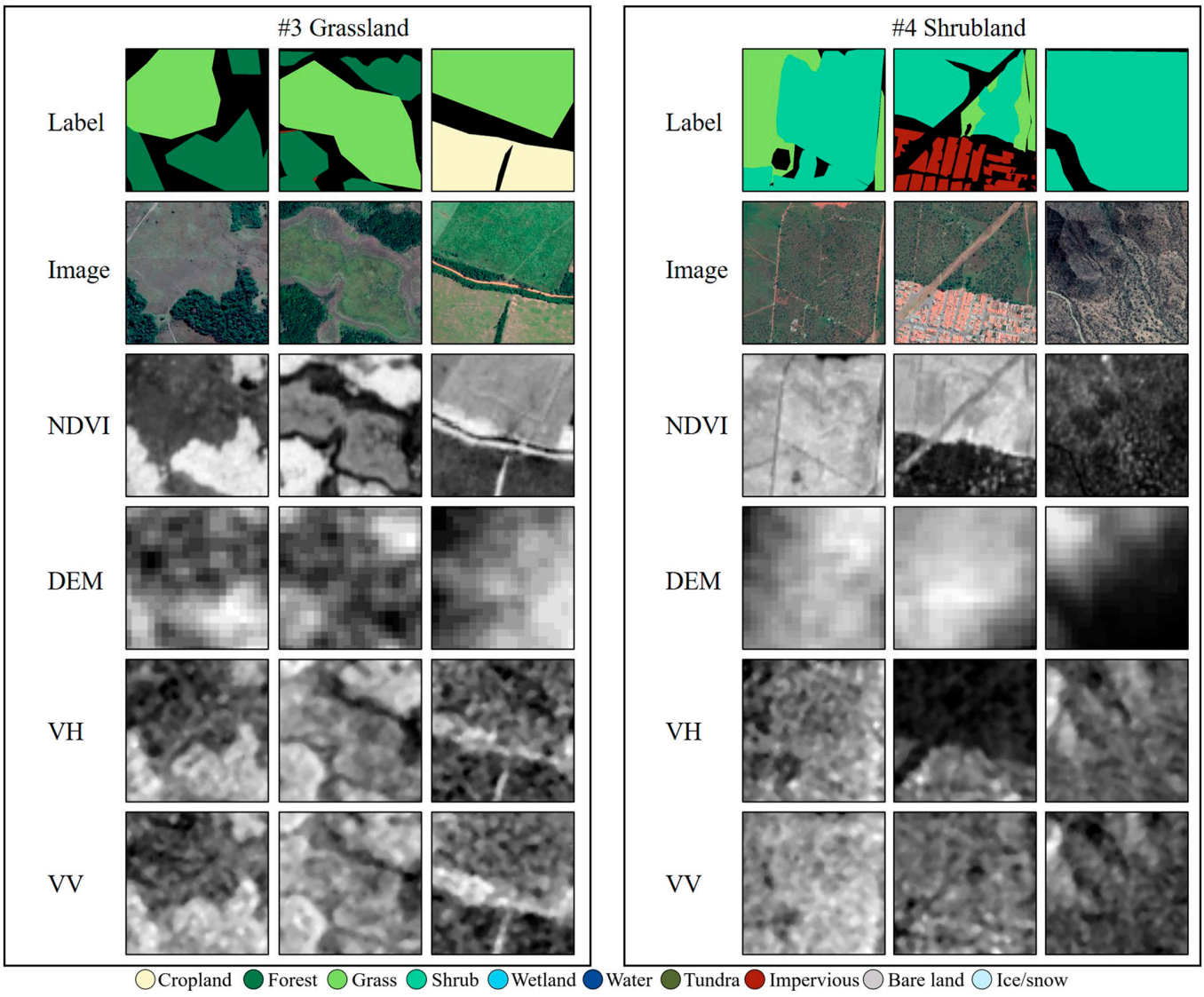

Fig. 6. Examples of the annotated patches of size 512 × 512 for grassland and shrubland.

the spatial location, which facilitates many variants for further improvement. For example, Ronneberger et al. [52] proposed the U-Net network to deepen the FCN structure and modified the decoder part of the FCN, to upsample the feature map size step by step, and fused with the corresponding feature map in the encoder part, which can gradually restore more spatial details, and the encoder–decoder forms a symmetric structure like shape "U." Considering the low efficiency of transmitting the low-level feature map from the encoder to the decoder part, SegNet [53] is proposed to transmit the indices of the pooling layers instead of the whole feature map to save memory. Deeping the network structure is also an alternative way for enlarging receptive field, but may confront the gradient vanishing or explosion problem. Therefore, He et al. [54] proposed residual network (ResNet) with skip connection strategy to iteratively transmit the low-level features with location information to the deeper layers. Huang et al. [55] proposed DenseNet based on the skip connection strategy, which concatenates the output of all the preview layers to enlarging receptive field while preserving the location information.

As for the practical application of FCN, Wei et al. [56] deployed the U-Net model on time-series for Sentinel-1 images for paddy rice mapping at Heilongjiang, Jilin, and Liaoning provinces of China. Yang et al. [57] combined U-Net and long short-term memory (LSTM) to extract the phenology feature of the paddy rice for inter-annual mapping. However, the loss of spatial location information during the pooling process is still difficult to restore. An alternative way for enlarging the receptive field without sacrificing location information is urgently needed.

Large convolution kernel for context extraction

Although pooling helps in increasing the receptive field, it decreases the spatial resolution, which is harmful for the spatial location prediction. In order to ensure a large receptive field while preserving spatial location information, researchers turn to focus on local property of convolution operation, and proposed the dilated convolution (also called atrous convolution) [58]. Dilated convolution uses very large kernel with hole to restrain the parameter increment caused by the enlargement of the kernel size, which can exponentially increase the receptive field. In this way, the spatial resolution can be preserved

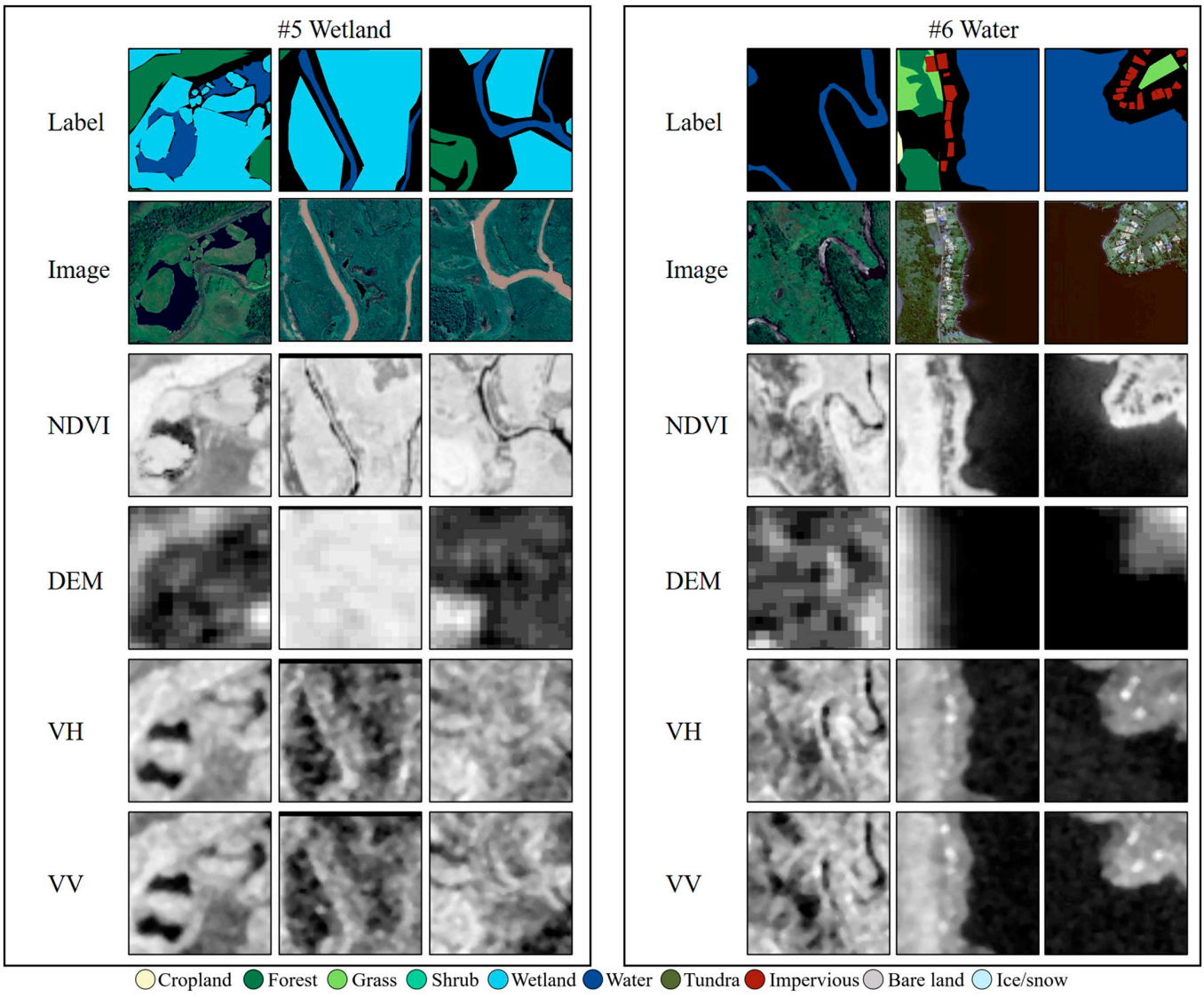

Fig. 7. Examples of the annotated patches of size 512 × 512 for wetland and water.

without pooling, while the richer context information can still be learned.

Specifically, Yu and Koltun [58] proposed a dilated convolution network (DilatedNet) to replace the pooling with dilated convolution in the deep layer and adopted different dilated rates (i.e., kernel size) to aggregate multiscale context. Considering that dilated rate should be manually predefined, He et al. [59] designed a strategy to adaptively adjust the dilated rate for different scene. Chen et al. [60] introduced the dilated convolution into deep convolutional network (DeepLabv1) for enlarging the receptive field and designed a fully connected conditional random field (CRF) during postprocessing to refine the location information. Peng et al. [61] proposed a global convolution network (GCN), used solid large kernel size to extract wider receptive field, and used separable convolution to separate the k × k large kernel size into the combination of 1 × k + k × 1 and k × 1 + 1 × k, which considerably reduces the amount of the parameters caused by large kernel. Furthermore, Panboonyuen et al. [62] deepened the GCN to accommodate medium-resolution remotely sensed images, and embedded the channel attention and the transfer learning module to alleviate the sample scarcity issue. Guo et al. [63] replaced the standard convolution with dilated convolution to enlarge the receptive field and embedded the graph-based segmentation and selective search [64] for high-resolution remote sensing image. However, since the size of the feature map remains large during feed forward, which occupies considerable computational memory, it is not suitable for high-resolution remote sensing image with large swath width.

Multiscale pyramid for context extraction

Due to the variation of altitude, imaging angle, and the land covers' distribution, land covers in remote sensing images exhibit multiscale effect. In order to accommodate the object of different scale, researchers attempt to aggregate the multiscale context information of the object in different stages of the semantic feature extracted by the encoder [65]. The aggregated multistage semantic features take into account different scales of the central object, making it possible to adapt to the recognition and positioning of objects under different scales.

For example, Zhao et al. [65] developed the pyramid scene parsing network (PSPNet), which uses multiscale pooling kernels with different sizes in parallel during encoder to extract

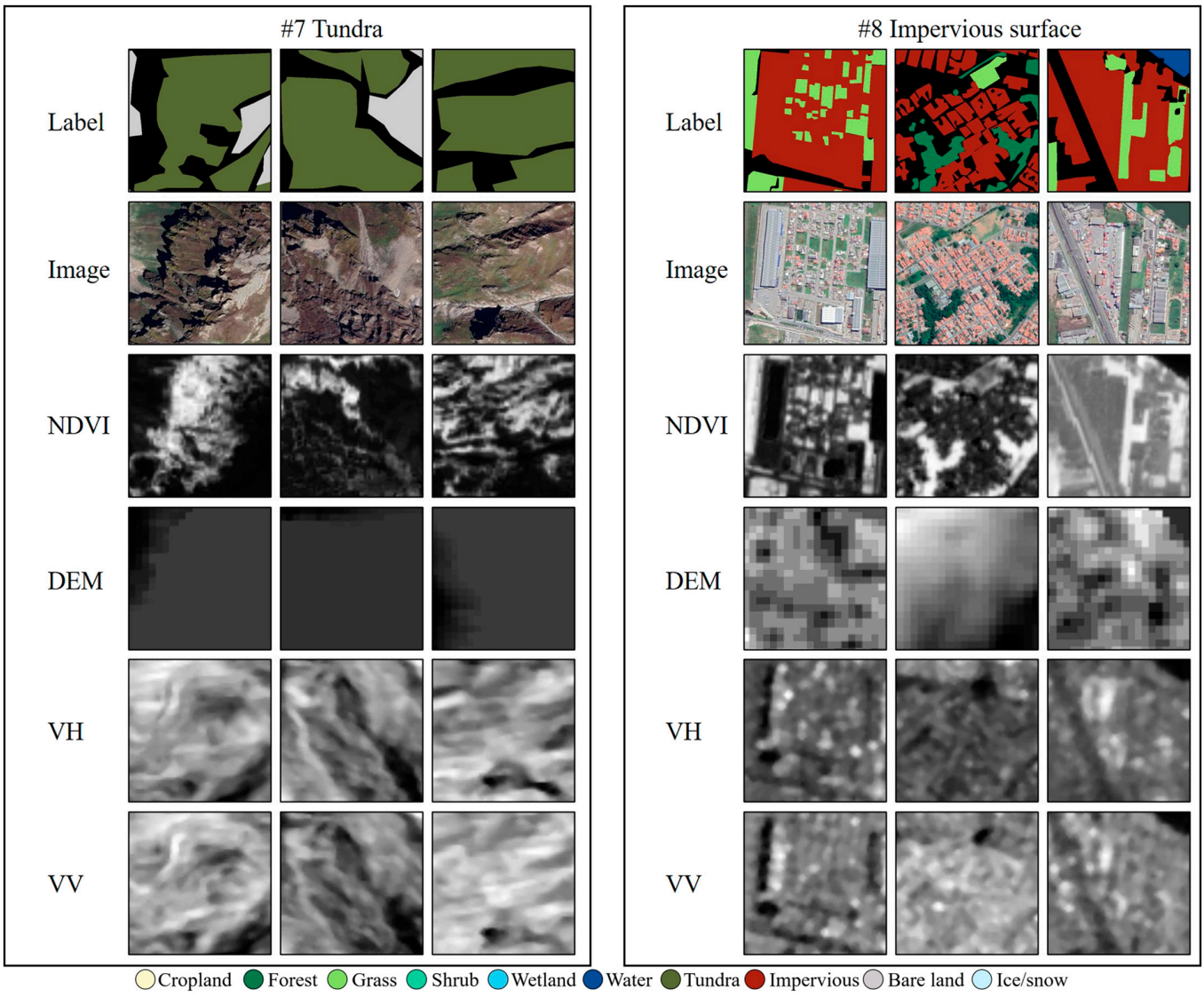

Fig. 8. Examples of the annotated patches of size 512 × 512 for tundra and impervious.

global and local scene semantic information at different stages. Lin et al. [66] proposed RefineNet to focus on the multiscale structure designment during decoder. It uses the residual connection to explicitly superimpose the output features of each downsampling layer of the encoder with the corresponding deconvolutional layer in the decoder, which is then fed to the residual pooling module to obtain the final feature map.

Furthermore, Chen et al. [67] further proposed the DeepLabv2 based on DeepLabv1, with the novel Atrous Spatial Pyramid Pooling (ASPP), to extract multiscale features from images. ASPP is composed of parallel dilated convolution kernels of different scales. It sequentially extracts features of different receptive field sizes from the input feature map and then fuses these multiscale features to obtain better semantic segmentation results. Chen et al. [68] further improved the ASPP module and proposed DeepLabv3, which introduced 1 × 1 convolution kernel and global pooling layer into ASPP to extract richer features, and introduced Batch Normalization (BN) [69] and other technologies to improve the convergence speed of network training. Chen et al. [70] further adopted an encoder–decoder structure based on DeepLabv3 and proposed DeepLabv3+, which fuses the low-level feature output with detailed location information by the encoder with the high-level feature output with semantic information by ASPP. Ji et al. [71] focused on the generalization of multiscale input images and proposed an SR-FCN network, which applied ASPP module to extract the multiscale context features, and evaluated the segmentation accuracy at different prediction scales to improve the performance of building segmentation on remote sensing image. Liu et al. [72] proposed a self-cascading semantic segmentation network (ScasNet) to optimize the segmentation performance of high-resolution remote sensing images through multiscale context aggregation, segmentation result refinement, and residual correction. Lin et al. [73] proposed a multi-scale context intertwining (MSCI) semantic segmentation network to use LSTM mechanism for bidirectionally fusing the multiscale features.

Edge-preserving strategy for location restoration

In order to enhance the segmented boundary of the land cover objects, an alternative way is to additionally design a boundary loss, which is optimized together with the segmentation loss

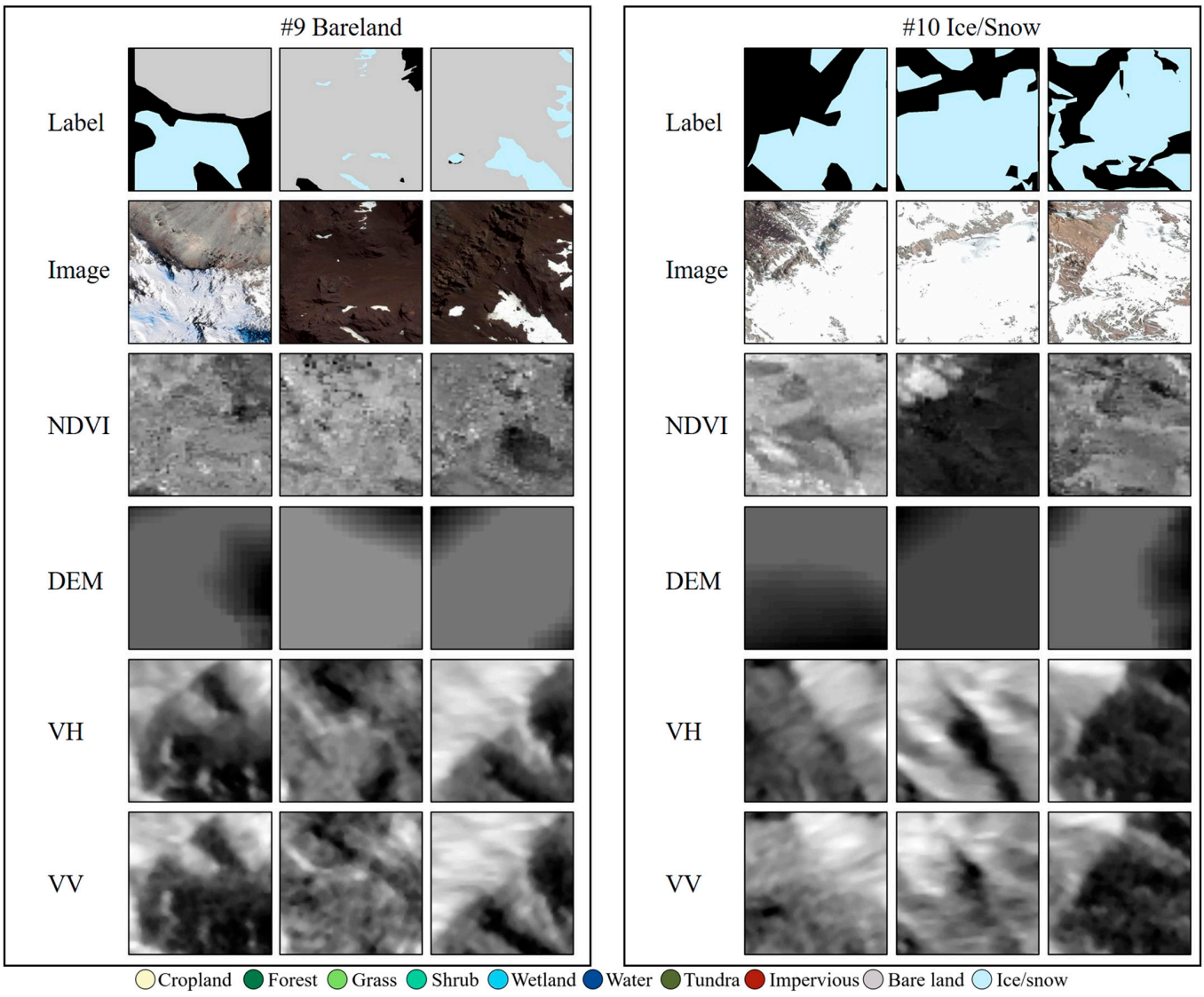

Fig. 9. Examples of the annotated patches of size 512 × 512 for bareland and ice.

during the network training. For example, Yuan [74] trained the network parameter by optimizing the distance loss from each pixel to the segmentation boundary and extracted the distribution of buildings in Washington DC, USA. Bischke et al. [75] applied the multitask learning strategy on the SegNet network, which combines the related tasks of segmentation and boundary prediction, to improve the accuracy of building extraction from high-resolution remote sensing images. Marmanis et al. [76] combined the edge information derived from the high resolution and the spectral information to enhance the building boundary segmentation. Mou and Zhu [77] proposed the auto-regressive recursive connection module to iteratively embed the low-level features with fine-grained location information and the high-level features with coarse semantic information layer by layer, which exhibits a performance promotion in the segmentation task of high-resolution image. Sun et al. [78] proposed the high-resolution network (HRNet), in which a multiresolution subnetwork in parallel is designed. Each of the high-to-low resolution representations in the subnetworks receives information from other parallel representations over and over, leading to rich high-resolution representations. HRNet was initially designed for human pose estimation and is now widely applied to remote sensing community [79]. Fu et al. [80] proposed Dual Attention Network (DANet) to adaptively integrate local features and global dependency features, and modeled the semantic interdependencies in spatial and channel dimensions to enhance both semantic and location information. Kirillov et al. [81] proposed the PointRend network that includes a progressive edge-guided deconvolution strategy to enable nonuniformed kernel specific at the low-confidence edge region and intensify the learning of the objects' boundary, under the inspiration of the image rendering process.

Transformer without pooling fashion

Recently, a new network architecture called Transformer has been boomed. It replaces the traditional convolution operation by full self-attention mechanism for feature representation, which can capture global context to compensate for the local connection limitation of convolution operation. Inspired

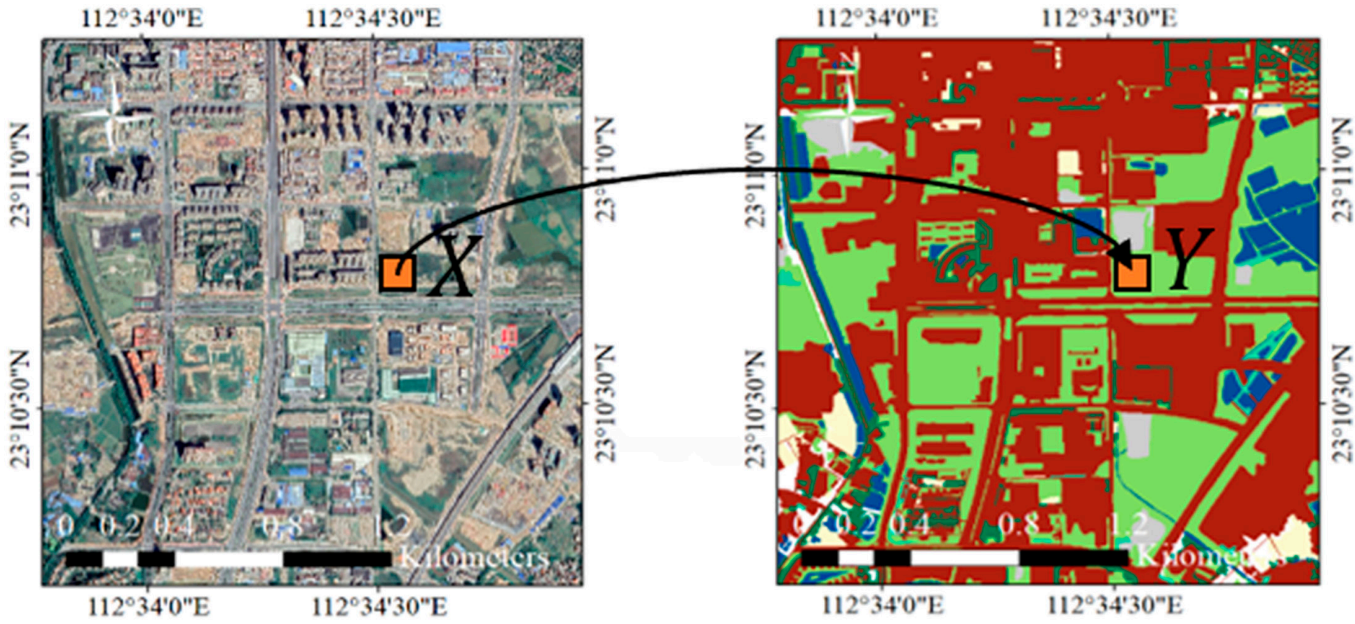

Fig. 10. Pixel-level understanding of the semantic segmentation task in Earth observation.

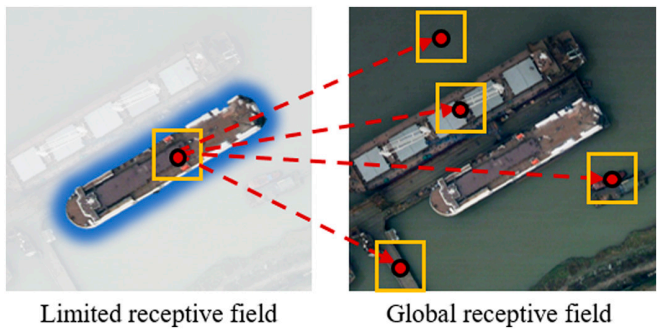

Fig. 11. Demonstration of how large receptive field with context information can help for the semantic recognition.

by the fact that words that are far apart in a sentence still have the strong correlation and contextual semantic association discovered in natural language processing community, the self-attention mechanism is introduced for long-distance dependency learning in the image scene.

Specifically, Dosovitskiy [82] proposed the primary Vision Transformer (ViT), which decomposed the image into multiple patches to analogy tokens, and the self-attention mechanism is taken to explore the long-distance dependency between tokens to extract contextual scene semantic information and help for better semantic segmentation. Notably, the size of the feature remains the same during the whole feed-forward process, which avoids the location information loss that confronted in the convolution fashion. Zheng [83] proposed the Semantic Segmentation Transformer (SETR) based on the encoder–decoder architecture, where the encoder consists of stacked multihead self-attention layers, and the decoder is composed of simple upsampling layers. Considering the limitation of the simple upsampling decoder, Xie [84] proposed the Trans2Seg to design a set of learnable prototypes as the query of Trans2Seg's Transformer decoder, and formulated semantic segmentation as a problem of dictionary look-up. Considering the fact that the feature size that remains the same across the whole network is not suitable for extracting multiscale representations, Wang et al. [85] designed the pyramid vision Transformer (PVT) in analogy to the convolution network fashion that progressively shrinks the resolution of the feature map to learn the multiscale features; however, the simple downsampling of the intermediate feature is harmful for long-distance dependency information. Liu [86] proposed a sliding window-based Transformer (Swin Transformer), which used sliding window to consider the patch tokens at the boundary of the scene to fully extract the global context during the feature size downsampling process. Considering that previous transformers only model the dependency among the patch tokens and the intrinsic structural information inside each patch token is ignored, Han et al. [87] proposed the Transformer-in-Transformer (TNT), which designed an inner Transformer block to model the dependency within each patch token, allowing the model to extract both global and local properties.

Benchmark Result

Experimental settings

In order to validate the effectiveness of the proposed Globe230k dataset, we select 10 representative deep learning-based sematic segmentation methods as our benchmark algorithms, which are first trained on the training set, and evaluated on the validation set during the training procedure. Finally, the best model is selected for prediction on the test set. Specifically, our selection includes 2 methods that use pooling for context extraction and deconvolution for location restoration, i.e., FCN [32] and ConvNeXt [88]. Three methods use multiscale pyramid for context extraction, i.e., DeepLabv3+ [70], PSPNet [65], and OCNet [89]. It should be noted that DeepLabv3+ also uses the atrous convolution for context extraction. Two methods use edge-preserving strategy for location restoration, i.e., HRNet [78] and PointRend [81]. Three methods use Transformer architecture, i.e., ViT [82], SegFormer [90], and SwinTransformer [86].

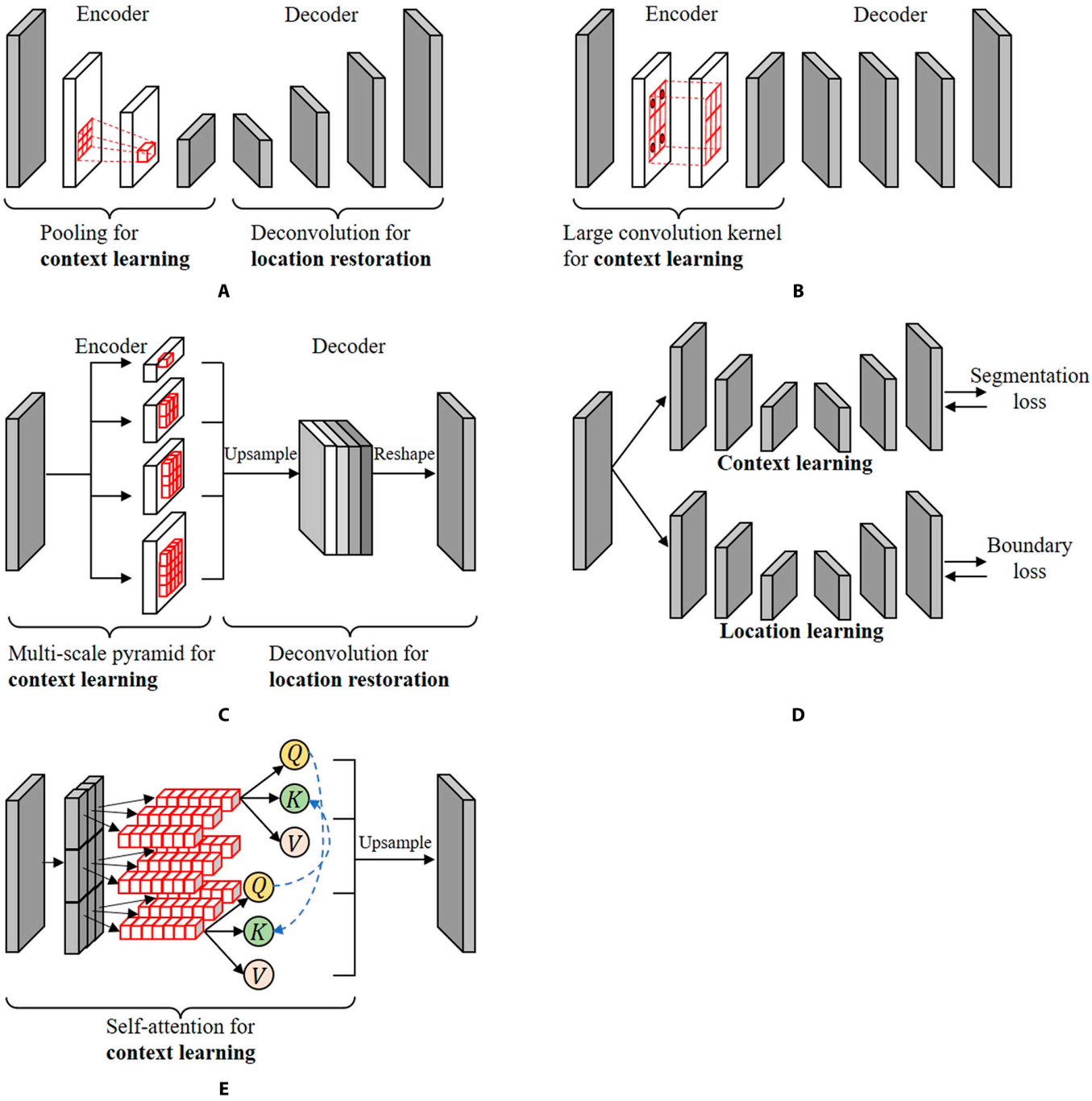

Fig.12. Systematic categorization of the state-of-the-art deep learning-based semantic segmentation methods. (A) Pooling for context learning. (B) Large convolution kernel for context learning. (C) Multi-scale pyramid for context learning. (D) Multi-branch structure for both context learning and location learning. (E) Self-attention for context learning.

All the experiments were conducted in PyTorch 1.9.0 and Python 3.7 on a server with 3 NVIDIA GeForce RTX3090 graphic processing unit (GPU) accelerators (with 24-GB GPU memory), and the implementation code adopted the mmsegmentation model library, which is available at https://github.com/open-mmlab/mmsegmentation. The number of epochs during the training process was uniformly set to 50, and the batch size was set to 16. Stochastic Gradient Descent (SGD) optimizer was applied with a momentum of 0.9 and a weight decay of 10^{-4}, and learning rate was initially set to 10^{-2} and reduced gradually under "poly" schedule with a power 0.9.

We used overall accuracy (OA) and mean of intersection over union of each class (mIoU) for overall performance evaluation, and used the precision accuracy (PA) of each class for class-wise performance evaluation.

Experimental results

The visual results of the selected representative methods are shown in Fig. 13, and the quantitative assessment is summarized in Table 2. Three conclusions can be drawn.

(a) The multiscale pyramid structure design can better promote the performance of the semantic segmentation on remote

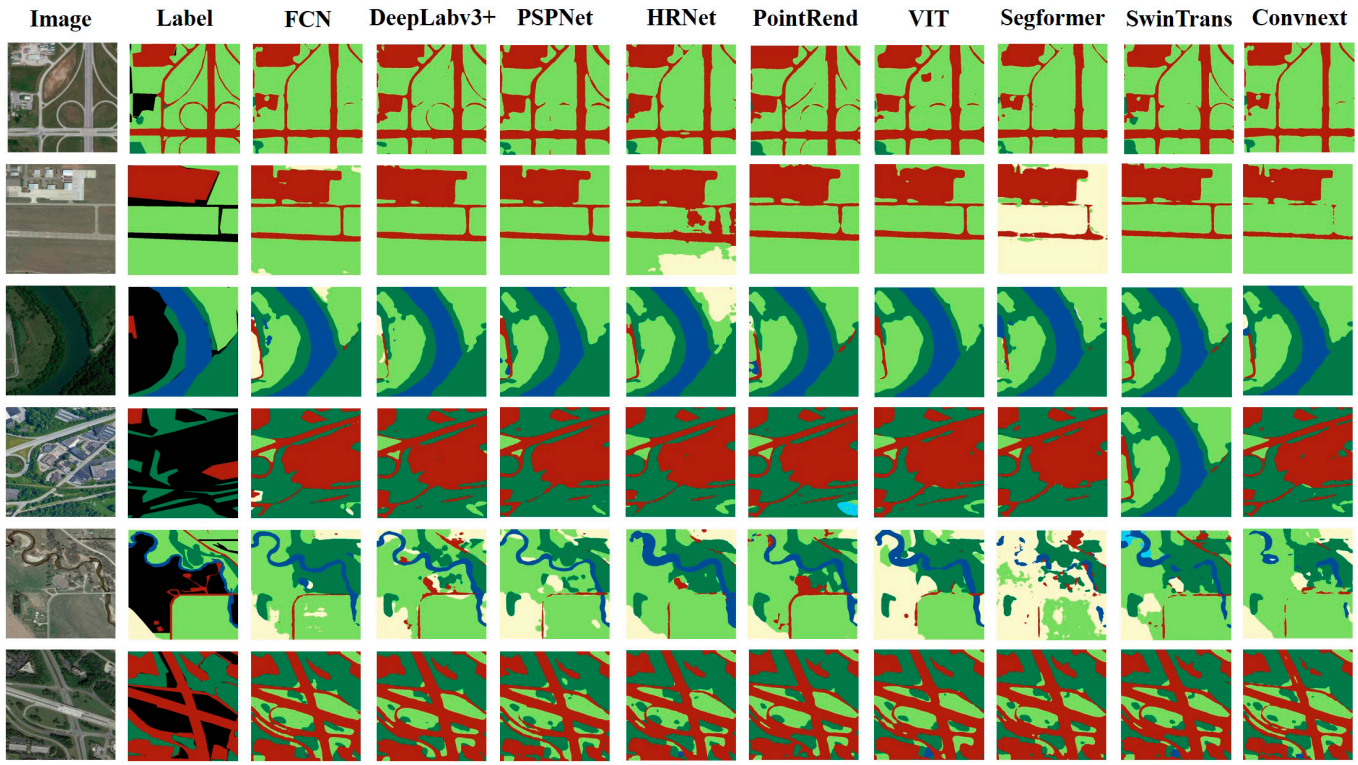

Fig. 13. Visual inspection of the benchmark semantic segmentation methods on the proposed Globe230k dataset.

Table 2. Quantitative assessment of the benchmark algorithms on the proposed Globe230k dataset.

Method	OA	mIoU	Category PA									
			Crop-land	Forest	Grass-land	Shrub-land	Wet-land	Water	Tundra	Imper-vious surface	Bare-land	Ice/ snow
FCN	89.12	67.97	90.73	95.37	71.90	70.20	68.23	88.34	0.30	94.03	59.59	96.26
Convnext	88.35	65.99	80.07	89.80	56.97	54.20	47.84	84.71	0.00	80.69	74.35	91.29
OCNet	87.45	70.12	89.98	89.71	75.34	57.58	49.62	82.67	27.91	80.11	74.78	91.45
DeepLabv3+	90.46	70.56	90.89	95.10	77.46	75.74	70.35	95.33	0.00	94.29	90.77	93.43
PSPNet	90.49	73.12	91.41	94.91	77.32	74.32	77.62	93.93	34.21	90.47	90.35	97.51
HRNet	88.20	65.78	91.94	96.18	59.86	61.36	61.24	91.66	0.00	91.36	93.18	93.70
PointRend	88.64	67.08	92.10	95.78	64.56	72.49	71.16	89.07	4.22	86.67	87.49	97.72
VIT	87.70	65.31	78.83	89.12	55.84	53.81	47.63	82.87	0.00	80.59	73.19	91.19
Segformer	86.69	62.64	92.65	94.64	55.14	56.13	41.96	92.95	0.00	92.36	90.05	95.27
SwinTransformer	90.90	75.72	89.42	96.52	79.33	71.73	77.28	94.99	42.64	93.06	90.74	98.06

sensing image, which is reflected by the fact that PSPNet and DeepLabv3+ achieve higher accuracy than most of the other methods, reaching 90.49% and 90.46% in OA and 73.12% and 70.56 in mIoU, respectively. The reason is that the multiscale pyramid design can take the multiscale effect derived from the sensor imaging or land cover's distribution characteristic into consideration, such as buildings with different sizes and orientations. Meanwhile, it can also aggregate large receptive field with different ranges, which further promote the recognition of land covers. Therefore, from this aspect, it also proved that our proposed Globe230k dataset can better express the distribution characteristics of the land cover objects, which make it possible to better evaluate the performance of the algorithms in terms of multiscale modeling. Besides, the advantage of the multiscale design is also significant in Transformer architecture by comparing the ViT and Swin Transformer,

since Swin Transformer improved the performance by 3.20% in terms of OA.

(b) From visual inspection, the PointRend performs well in detailed location restoration, e.g., the elongated viaduct and road are more coherent than the results in other methods, which owns to the designment of the boundary intensification mechanism during the progressive upsampling in the decoder part and the auxiliary boundary loss. Therefore, it also proves that the Globe230k dataset is fine-grained enough to evaluate the reconstruction ability of different algorithms.

(c) By comparing the best convolutional-based method (PSPNet) and the best Transformer-based method (Swin Transformer), we can observe that the self-attention performs better than convolution operation in semantic recognition due to its global receptive field for context learning, but pure self-attention design is unfriendly in local information prediction (as can be seen in the result of ViT). Thus, combining self-attention and convolution is a promising way for a better performance. From this point of view, the Globe230k dataset can reflect the global characteristic of the geospatial distribution.

Noteworthy is that tundra is a low-growing vegetation inhabited in extremely cold temperatures, which is typically found in the Earth's northernmost regions, such as the Arctic and high-altitude areas, and it is very similar to grassland class. Therefore, the accuracy of tundra is usually compromised, with large amount of tundra mistakenly identified as grassland. Another reason is that there is almost no tundra in China. It is difficult for us to field survey and verify the annotations, and thus, only a few tundra samples are in the Globe230k dataset, resulting in an unsatisfied tundra result.

Conclusion

This study is aimed to facilitate the generalization ability and spatial resolving ability of the semantic segmentation networks for global land cover mapping activity by creating the large-scale remote sensing image semantic segmentation annotation dataset Globe230k based on our proposed annotation principal, and the numerous time and effort spend on it by experts, teachers, and students. Besides, we also make a systematic review of the recent progress of the semantic segmentation algorithms and their implementations in remote sensing community, and also the benchmarking land cover datasets in Earth observation. The representative and state-of-the-art semantic segmentation algorithms are selected and evaluated by the proposed dataset, and the experimental results can provide an insight into algorithm designment and global land cover mapping activity. We believe that the Globe230k dataset could support further Earth observation research and provide a brand-new insight into the global land cover dynamic monitoring.

Acknowledgments

Funding: This work was supported by the National Key Research and Development Program of China (2022YFB3903402) and the National Natural Science Foundation of China (42222106, 61976234, and 42201340). **Author contributions**: Q.S.: Conceptualization, resources, data curation, writing—review and editing, and funding acquisition. D.H.: Methodology, software, validation, formal analysis, writing—original draft, and funding acquisition. Z.L.: Resources and data curation. X.L.: Supervision and funding acquisition. J.X.: Software and validation. **Competing interests**: The authors declare that they have no competing interests.

Data Availability

The Globe230k dataset is available at (https://doi.org/10.5281/zenodo.8429200).

References

1. Liu X, Huang Y, Xu X, Li X, Li X, Ciais P, Lin P, Gong K, Ziegler AD, Chen A, et al. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. *Nat Sustain*. 2020;3(7):564–570.
2. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, et al. High-resolution global maps of 21st-century forest cover change. *Science*. 2013;342(6160):850–853.
3. Li L, Chen Y, Yu X, Liu R, Huang C. Sub-pixel flood inundation mapping from multispectral remotely sensed images based on discrete particle swarm optimization. *ISPRS J Photogramm Remote Sens*. 2015;101:10–21.
4. Li D, Liao W, Rigden AJ, Liu X, Wang D, Malyshev S, Shevliakova E. Urban heat island: Aerodynamics or imperviousness? *Sci Adv*. 2019;5(4):eaau4299.
5. Feddema JJ, Oleson KW, Bonan GB, Mearns LO, Buja LE, Meehl GA, Washington WM. The importance of land-cover change in simulating future climates. *Science*. 2005;310(5754):1674–1678.
6. Beusch L, Gudmundsson L, Seneviratne SI. Crossbreeding CMIP6 earth system models with an emulator for regionally optimized land temperature projections. *Geophys Res Lett*. 2020;47(15):GL086812.
7. Loveland TR, Reed BC, Brown JF, Ohlen DO, Zhu Z, Yang L, Merchant JW. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *Int J Remote Sens*. 2000;21(6-7):1303–1330.
8. Friedl MA, Sulla-Menashe D, Tan B, Schneider A, Ramankutty N, Sibley A, Huang X. MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sens Environ*. 2010;114(1):168–182.
9. Li W, MacBean N, Ciais P, Defourny P, Lamarche C, Bontemps S, Houghton RA, Peng S. Gross and net land cover changes in the main plant functional types derived from the annual ESA CCI land cover maps (1992–2015). *Earth Syst Sci Data*. 2018;10(1):219–234.
10. Defourny P. GLOBCOVER: A 300 m global land cover product for 2005 using Envisat MERIS time series. Paper presented at: Proceedings of ISPRS Commission VII Mid-Term Symposium: Remote Sensing: From Pixels to Processes; 2006; Enschede (NL).
11. Chen J, Chen J. GlobeLand30: Operational global land cover mapping and big-data analysis. *Sci China Earth Sci*. 2018;61(10):1533–1534.
12. Gong P, Wang J, Yu L, Zhao Y, Zhao Y, Liang L, Niu Z, Huang X, Fu H, Liu S, et al. Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. *Int J Remote Sens*. 2013;34(7):2607–2654.
13. Karra K, Kontgis C, Statman-Weil Z, Mazzariello JC, Mathis M, Brumby SP. Global land use/land cover with Sentinel 2 and deep learning. Paper presented at: 2021 IEEE International

Geoscience and Remote Sensing Symposium IGARSS: 2021 Jul 11–16; Brussels, Belgium.
14. Zhang X, Liu L, Chen X, Gao Y, Xie S, Mi J. GLC_FCS30: Global land-cover product with fine classification system at 30 m using time-series Landsat imagery. *Earth Syst Sci Data*. 2021;13(6):2753–2776.
15. Zanaga D. ESA WorldCover 10 m 2021 v200; 2022.
16. Blaschke T, Lang S, Lorup E, Strobl J, Zeil P. Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications. *Geoscience*. 2000;2(1995):555–570.
17. Duro DC, Franklin SE, Dubé MG. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens Environ*. 2012;118:259–272.
18. Foody GM. Land cover classification by an artificial neural network with ancillary information. *Int J Geogr Inf Syst*. 1995;9(5):527–542.
19. Hu T, Huang X, Li J, Zhang L. A novel co-training approach for urban land cover mapping with unclear Landsat time series imagery. *Remote Sens Environ*. 2018;217:144–157.
20. Myint SW, Gober P, Brazel A, Grossman-Clarke S, Weng Q. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sens Environ*. 2011;115(5):1145–1161.
21. Yan G, Mas JF, Maathuis B, Xiangmin Z, Van Dijk P. Comparison of pixel-based and object-oriented image classification approaches—A case study in a coal fire area, Wuda, Inner Mongolia, China. *Int J Remote Sens*. 2006;27(18):4039–4055.
22. Abou El-Magd I, Tanton TW. Improvements in land use mapping for irrigated agriculture from satellite sensor data using a multi-stage maximum likelihood classification. *Inter J Remote Sens*. 2003;24(21):4197–4206.
23. Peña JM, Gutiérrez PA, Hervás-Martínez C, Six J, Plant RE, López-Granados F. Object-based image classification of summer crops with machine learning methods. *Remote Sens*. 2014;6(6):5019–5041.
24. Löw F, Michel U, Dech S, Conrad C. Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines. *ISPRS J. Photogramm Remote Sens*. 2013;85:102–119.
25. Peña-Barragán JM, Ngugi MK, Plant RE, Six J. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. *Remote Sens Environ*. 2011;115(6):1301–1316.
26. Watts J, Lawrence R. Merging random forest classification with an object-oriented approach for analysis of agricultural lands. *Biomed Signal Sens*. 2008;37(B7):1.
27. Li SZ. *Markov random field modeling in image analysis*. London: Springer Science & Business Media; 2009.
28. Cruz-Ramírez M, Hervás-Martínez C, Jurado-Expósito M, López-Granados F. A multi-objective neural network based method for cover crop identification from remote sensed data. *Expert Syst Appl*. 2012;39(11):10038–10048.
29. Kemker R, Salvaggio C, Kanan C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J Photogramm Remote Sens*. 2018;145:60–77.
30. Zhang J, Lin S, Ding L, Bruzzone L. Multi-scale context aggregation for semantic segmentation of remote sensing images. *Remote Sens*. 2020;12(4):701.
31. Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J Photogramm Remote Sens*. 2019;152:166–177.
32. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2015; UC Berkeley, USA.
33. Yuan X, Sarma V. Automatic urban water-body detection and segmentation from sparse ALSM data via spatially constrained model-driven clustering. *IEEE Geosci Remote Sens Lett*. 2010;8(1):73–77.
34. Yang S, Chen Q, Yuan X, Liu X. Adaptive coherency matrix estimation for polarimetric SAR imagery based on local heterogeneity coefficients. *IEEE Trans Geosci Remote Sens*. 2016;54(11):6732–6745.
35. Kussul N, Lavreniuk M, Skakun S, Shelestov A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci Remote Sens Lett*. 2017; 14(5):778–782.
36. Dechesne C, Mallet C, Le Bris A, Gouet-Brunet V. Semantic segmentation of forest stands of pure species as a global optimization problem. *ISPRS Ann Photogramm Remote Sens Spatial Inf Sci*. 2017;4:141–148.
37. Rottensteiner F, Sohn G, Jung J, Gerke M, Baillard C, Benitez S, Breitkopf U. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann Photogramm Remote Sens Spatial Infor Sci*. 2012;1(1):293–298.
38. Volpi M, Ferrari V. Semantic segmentation of urban scenes by learning local class interactions. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops: 2015 Jun 7–12; Boston, MA, USA.
39. Meraner A, Ebel P, Zhu XX, Schmitt M. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *ISPRS J Photogramm Remote Sens*. 2020;166:333–346.
40. Jia K, Liang S, Zhang L, Wei X, Yao Y, Xie X. Forest cover classification using Landsat ETM+ data and time series MODIS NDVI data. *Int J Appl Earth Obs Geoinf*. 2014;33:32–38.
41. Bahadur KKC. Improving Landsat and IRS image classification: Evaluation of unsupervised and supervised classification through band ratios and DEM in a mountainous landscape in Nepal. *Remote Sens*. 2009;1(4):1257–1272.
42. Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S, Houlsby N. Big transfer (bit): General visual representation learning. Paper presented at: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16: 2020 Aug 23; Berlin, Heidelberg.
43. Shao Z, Yang K, Zhou W. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sens*. 2018;10(6):964.
44. Wang J, Zheng Z, Ma A, Lu X, Zhong Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation, ArXiv 2021. arXiv preprint arXiv:2110.08733.
45. Li J, Huang X, Tu L. WHU-OHS: A benchmark dataset for large-scale Hersepctral image classification. *Int J Appl Earth Obs Geoinf*. 2022;113:103022.
46. Tong X, Xia GS, Lu Q, Shen H, Li S, You S, Zhang L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens Environ*. 2020;237:111322.
47. International Society for Photogrammetry and Remote Sensing 2D Semantic Labeling Challenge; https://www.isprs.

org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx.
48. Kotaridis I, Lazaridou M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS J Photogramm Remote Sens*. 2021;173:309–322.
49. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J Photogramm Remote Sens*. 2020;162:94–114.
50. Sultana F, Sufian A, Dutta P. Evolution of image segmentation using deep convolutional neural network: A survey. *Knowl-Based Syst*. 2020;201:106062.
51. Ding L, Zhang J, Bruzzone L. Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture. *IEEE Trans Geosci Remote Sens*. 2020;58(8):5367–5376.
52. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation.In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 2015, Proceedings, Part III 18 (pp. 234-241).
53. Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2007;39(12):2481–2495.
54. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV.
55. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, Hawaii, USA. p. 4700-4708.
56. Wei P, Chai D, Huang R, Peng D, Lin T, Sha J, Sun W, Huang J. Rice mapping based on Sentinel-1 images using the coupling of prior knowledge and deep semantic segmentation network: A case study in Northeast China from 2019 to 2021. *Int J Appl Earth Obs Geoinf*. 2022;112:102948.
57. Yang L, Huang R, Huang J, Lin T, Wang L, Mijiti R, Wei P, Tang C, Shao J, Li Q, et al. Semantic segmentation based on temporal features: Learning of temporal–spatial information from time-series SAR images for paddy rice mapping. *IEEE Trans Geosci Remote Sens*. 2021;60:1–16.
58. Yu F, Koltun V, Multi-scale context aggregation by dilated convolutions, ArXiv 2015. arXiv preprint arXiv:1511.07122.
59. He Y, Keuper M, Schiele B, Fritz M. *Learning dilation factors for semantic segmentation of street scenes*. Basel (Switzerland): Springer; 2017. p. 41-51.
60. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL, Semantic image segmentation with deep convolutional nets and fully connected crfs. ArXiv 2014. arXiv preprint arXiv:1412.7062.
61. Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters—Improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, Hawaii, USA. p. 4353-4361.
62. Panboonyuen T, Jitkajornwanich K, Lawawirojwong S, Srestasathiern P, Vateekul P. Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning. *Remote Sens*. 2019;11(1):83.
63. Guo R, Liu J, Li N, Liu S, Chen F, Cheng B, Duan J, Li X, Ma C. Pixel-wise classification method for high resolution remote sensing imagery using deep neural networks. *ISPRS Int J Geo Inf*. 2018;7(3):110.
64. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW. Selective search for object recognition. *Int J Comput Vis*. 2013;104:154–171.
65. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, Hawaii, USA. p. 2881-2890.
66. Lin G, Milan A, Shen C, Reid I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; Honolulu, Hawaii, USA. p. 1925-1934.
67. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans Pattern Anal Mach Intell*. 2017;40(4):834–848.
68. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking Atrous convolution for semantic image segmentation. CoRR. 2017;abs/1706.05587.
69. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR. 2015;abs/1502.03167.
70. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Paper presented at: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8; Berlin, Heidelberg.
71. Ji S, Wei S, Lu M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int J Remote Sens*. 2019;40(9):3308–3322.
72. Liu Y, Fan B, Wang L, Bai J, Xiang S, Pan C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J Photogramm Remote Sens*. 2018;145:78–95.
73. Lin D, Ji Y, Lischinski D, Cohen-Or D, Huang H. Multi-scale context intertwining for semantic segmentation. Paper presented at: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8; Berlin, Heidelberg.
74. Yuan J. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;40(11):2793–2798.
75. Bischke B, Helber P, Folz J, Borth D, Dengel A. Multi-task learning for segmentation of building footprints with deep neural networks. Paper presented at: 26th IEEE International Conference on Image Processing (ICIP): 2019 Sep 22–25; Taipei, Taiwan.
76. Marmanis D, Schindler K, Wegner JD, Galliani S, Datcu M, Stilla U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J Photogramm Remote Sens*. 2018;135:158–172.
77. Mou L, Zhu XX. RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. ArXiv 2018. arXiv preprint arXiv:1805.02091.
78. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA.

79. Seong S, Choi J. Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates. *Remote Sens*. 2021;13(16):3087.
80. Fu J. Dual attention network for scene segmentation. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA.
81. Kirillov A, Wu Y, He K, Girshick R. Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 9799-9808, Virtual.
82. Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv 2020. arXiv preprint arXiv:2010.11929.
83. Zheng S. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 6881-6890.
84. Xie E. Segmenting transparent object in the wild with transformer. ArXiv 2021. arXiv preprint arXiv:2101.08461.
85. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Paper presented at: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada.
86. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 10012-10022.
87. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. *Adv Neural Inf Proces Syst*. 2021;34:15908–15919.
88. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA.
89. Yuan Y, Huang L, Guo J, Zhang C, Chen X, WangJ. Ocnet: Object context network for scene parsing. ArXiv 2018. arXiv preprint arXiv:1809.00916.
90. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv Neural Inf Proces Syst*. 2021;34:12077–12090.
91. Bahmanyar R, Espinoza-Molina D, Datcu M. Multisensor earth observation image classification based on a multimodal latent Dirichlet allocation model. *IEEE Geosci Remote Sens Lett*. 2018;15(3):459–463.
92. Pilkington N, Svetlichnaya S, Holmes T. DroneDeploy's aerial segmentation benchmark; 2019.
93. Ji S, Zhang Z, Zhang C, Wei S, Lu M, Duan Y. Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images. *Int J Remote Sens*. 2020;41(8):3162–3174.
94. Semantic Drone Dataset, http://dronedataset.icg.tugraz.at.
95. Castillo-Navarro J, Le Saux B, Boulch A, Audebert N, Lefèvre S. Semi-supervised semantic segmentation in earth observation: The MiniFrance suite, dataset analysis and multi-task network study. *Mach Learn*. 2021;1–36.
96. Demir I, Koperski K, Lindenbaum D, Pang G, Huang J, et al. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018; Salt Lake City, USA. p. 172-181.
97. Azimi SM, Henry C, Sommer L, Schumann A, Vig E. SkyScapes—Fine-grained semantic understanding of aerial scenes. Paper presented at: IEEE/CVF International Conference on Computer Vision (ICCV); 2019; Seoul, South Korea.