

# Seeing Beyond the Patch: Scale-Adaptive Semantic Segmentation of High-resolution Remote Sensing Imagery based on Reinforcement Learning

Yinhe Liu, Sunan Shi, Junjue Wang, Yanfei Zhong\*

Wuhan University, Wuhan, China

{liuyinhe, shisunan, kingdrone, zhongyanfei}@whu.edu.cn

## Abstract

In remote sensing imagery analysis, patch-based methods have limitations in capturing information beyond the sliding window. This shortcoming poses a significant challenge in processing complex and variable geo-objects, which results in semantic inconsistency in segmentation results. To address this challenge, we propose a dynamic scale perception framework, named GeoAgent, which adaptively captures appropriate scale context information outside the image patch based on the different geo-objects. In GeoAgent, each image patch's states are represented by a global thumbnail and a location mask. The global thumbnail provides context beyond the patch, and the location mask guides the perceived spatial relationships. The scale-selection actions are performed through a Scale Control Agent (SCA). A feature indexing module is proposed to enhance the ability of the agent to distinguish the current image patch's location. The action switches the patch scale and context branch of a dual-branch segmentation network that extracts and fuses the features of multi-scale patches. The GeoAgent adjusts the network parameters to perform the appropriate scale-selection action based on the reward received for the selected scale. The experimental results, using two publicly available datasets and our newly constructed dataset WUSU, demonstrate that GeoAgent outperforms previous segmentation methods, particularly for large-scale mapping applications.

## 1. Introduction

The acquisition of high spatial resolution (HSR) remote sensing images has become possible with the advancement of sensors [4]. These images enable the extraction of geospatial objects, such as buildings, roads, and aircraft, through image interpretation [47, 45, 22, 53]. Furthermore, HSR remote sensing images facilitate the determination of more semantic land use/ land cover (LULC) categories, including industrial land, cropland, and residential land [8]. Such LULC mapping is critical in urban management, plan-

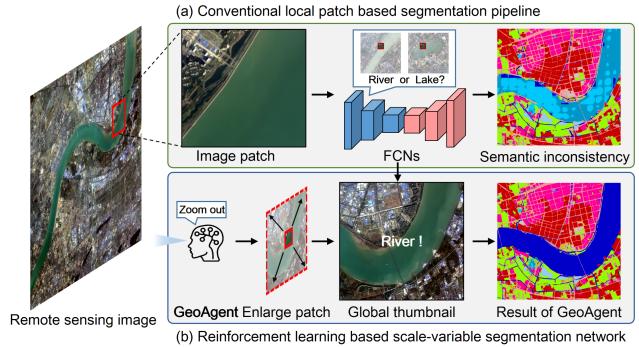


Figure 1: Illustration of the scale problem for HSR remote sensing image LULC segmentation. The sliding window  $\square$  is enlarged for a better view. Legend: ■ river, ■ lake, ■ pond.

ning, and ecological conservation [3, 16, 30].

The task of LULC mapping requires the semantic categorization of each pixel, also known as semantic segmentation. Recently, Fully Convolutional Networks (FCNs) [25] have achieved significant progress for natural images in computer vision [43, 49, 17, 33, 50, 5]. Deep learning techniques have also shown success in the automatic interpretation of HSR remote sensing images [15, 10, 51, 21, 29].

In contrast to conventional ground-based images, HSR remote sensing images have a large length and width that can encompass thousands of objects in a single image. However, distinguishing LULC (Land Use and Land Cover) geo-objects that vary greatly at different scales using a fixed, finite size image patch is challenging. Although Fully Convolutional Networks (FCNs) can theoretically receive input of arbitrary size, when processing HSR remote sensing images, a sliding window (SW) of a specific size is often used to generate image patches [41, 40, 9]. In small patches, different LULC classes may consist of the same objects with similar patterns, making them almost indistinguishable [52, 48]. For instance, the same water may appear in rivers, lakes, and ponds, and the same dense housing may belong to rural or urban residential land. Moreover, some cultivated land and meadows may have similar

\*corresponding author.

soils [37, 38]. Regardless of the multi-scale module’s design [5, 19, 23, 50, 13, 44, 24], the model’s receptive field cannot surpass the image patch’s size [1]. Consequently, the commonly used Deeplabv3+ [5] model’s mapping results exhibit an apparent grid effect due to the gap between the sliding window’s limited size and the river’s enormous scale, as illustrated in Figure 1.

The accurate segmentation of land use and land cover (LULC) in high spatial resolution (HSR) remote sensing images poses a challenge due to the significant scale variance and limited receptive field of the network by patch size. Various models aim to address this issue by tackling the multiscale problem from the perspective of input data [6, 11, 7]. While incorporating contextual information can effectively resolve the problem of identifying rivers and water bodies within limited image patches, different LULC geo-objects necessitate different scales of image patches. For instance, local image patches or slightly larger scales may suffice for identifying a pit pond, whereas lakes and continuous rivers may require larger scales. An inappropriate scale of receptive field may have an adverse impact on the segmentation.

The supervised learning methods have difficulty in adaptively determining the appropriate patch scales, as it requires manual labeling of images that correspond to the appropriate context scale. In contrast, reinforcement learning (RL) algorithms operate through interactions between the *agent* and the *environment*, receiving rewards for *actions* and using that information to update model parameters, without requiring labeled data [36]. RL techniques have achieved success in various areas, including board games, video games, and mathematics [27, 35, 14]. Thus, applying deep RL techniques to LULC segmentation networks may allow for the adaptive adjustment of the receptive field scale based on image features, dynamically controlling the data flow of HSR remote sensing images. Therefore, we propose a scale-adaptive segmentation network, called GeoAgent to dynamically select appropriate patch scale based on different geo-objects. A HSR remote sensing segmentation dataset WUSU is constructed for experiment. The main contributions include:

(1) A RL-based HSR imagery segmentation method is proposed. In GeoAgent, the HSR remote sensing images are represented by different *states*. GeoAgent observes different *states* and performs a series of scale-selection *actions* to control the area of each patch, obtaining the final global mapping *reward* based on the selected scales.

(2) A Scale Control Agent (SCA) is proposed to dynamically controls the scale of the GeoAgent. The SCA contains two parts: an ‘actor’ network that learns a policy to map the *states* to a probability distribution of *actions* and a ‘critic’ network that evaluates the value of each *action* using an advantage function. A feature indexing module is proposed to

enhance the *agent*’s ability to distinguish the location of the current image patch.

(3) GeoAgent use a dual-branch segmentation network for multi-scale image patches segmentation. The input scale and context branch are switched under the control of the *actions* selected by SCA. The network processes the multi-scale context patches and local scale patches separately, and the multi-scale features are aggregated based on the geo-location information to generate the final multi-scale segmentation result.

## 2. Related Works

### 2.1. Multi-Scale Segmentation Networks

There have been several techniques to solve multi-scale problems, ranging from multi-scale feature ensemble (e.g. Deeplab [5], FPN [19]) to context information preservation (e.g. ParseNet [23], PSPNet [50]). Through the self-attention mechanism, it is able to capture long-distance relationships in images [13, 44, 24]. But no matter how the multi-scale module is designed, the receptive field of the model cannot break through the size of the sliding window.

Therefore some models try to solve the multi-scale problem from the perspective of input data. [32] proposed a saliency-based distortion layer for convolutional neural networks to improve the spatial sampling of input data with the condition of the given task. Some methods fetch larger scale image outside the patch to provide spatial context. [6] proposed the GLNet, which consists of a global branch and a local branch, which take the downsampled global image and the full-resolution cropped image as input, respectively, effectively preserving the detail information and the global context information. [11] propose a wide-context network (WiCoNet), a context transformer is introduced to extract the downsampled global context image. On the other hand, some approaches address the problem by leveraging multi-scale processing. [7] proposed CascadePSP, which first performs segmentation on the downsampled coarse resolution images, and then uses the original resolution images to refine the segmentation results. [18] proposed multi-scale framework MagNet that uses multiple processing stages to resolve local ambiguity by examining the image at various magnification levels and propagating coarse-to-fine information. However, some of the above methods do not break the limits of image patches [32], or introduced a fixed scale context image[6, 11] and complex processes [18, 7] that do not allow the model adaptively choose appropriate scale to zoom the sliding window based on different objects.

### 2.2. DRL-based Dynamic Networks

Deep Reinforcement Learning (DRL) is able to autonomously make decisions about the following *action* by observing the *state* of environment, which in turn affects the environment, and updating the model parameters by re-

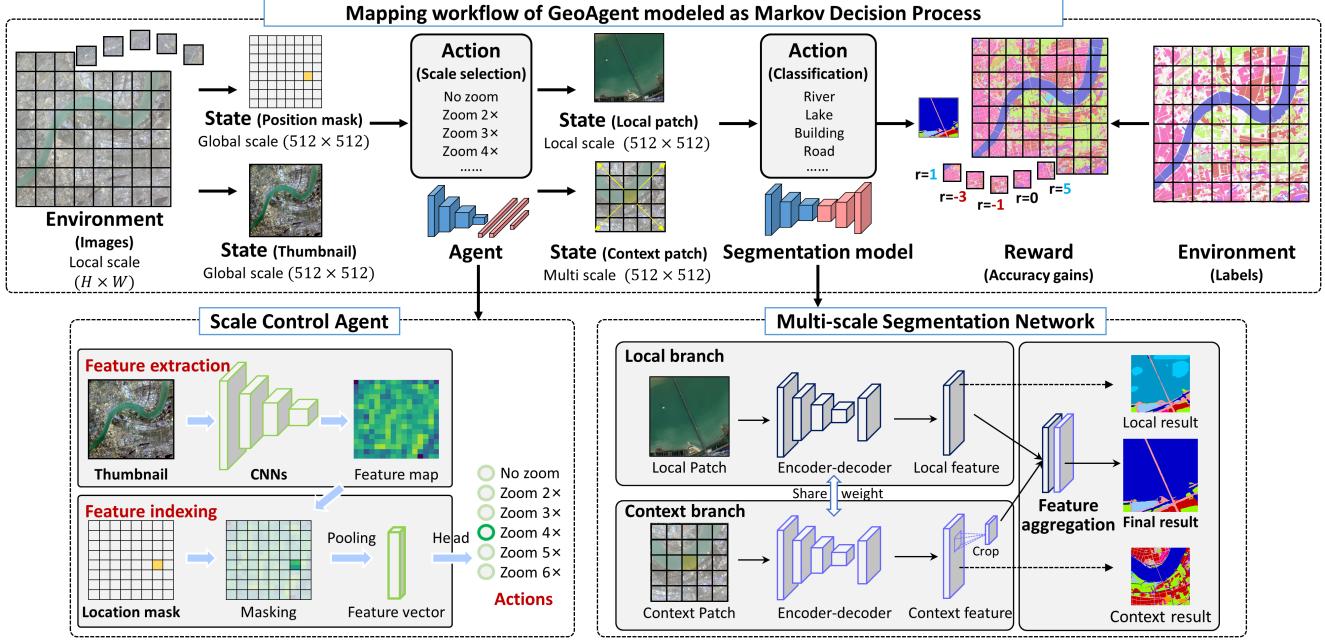


Figure 2: The overall flowchart of the proposed reinforcement learning based scale-adaptive segmentation network GeoAgent.

ceiving the environment’s *reward* for the *action*. By introducing the DRL approach, CNNs can be more dynamic and able to automatically change the network structure based on the input or influence the input to the network. [46] uses reinforcement learning to train a policy network for double *reward*. During training and testing time, the *agent* uses the input image as a *state* to try to utilize the least number of blocks while maintaining recognition accuracy. [39] proposed a Patch-Drop algorithm to selectively use high-resolution satellite data when necessary. A policy network is utilized to take low-resolution remote sensing images as *state* to decide whether to sample high-resolution remote sensing images as additional inputs. [2] proposed a reinforcement learning setup to acquire high-resolution satellite images conditionally. A cost-aware *reward* function was designed to reduce satellite image acquisition costs while maintaining accuracy. [12] proposed RAZN to achieve accurate and fast prediction of breast cancer segmentation in whole-slide images by learning a policy network to decide whether zooming is required in a given region of interest, motivated by the zoom-in operation of a pathologist using a digital microscope. To the authors’ knowledge, there has yet to be work using RL to solve the problem of limited patch size in remote sensing imagery segmentation task, so the scale-variable network GeoAgent is proposed.

### 3. Methodology

#### 3.1. Overview

The GeoAgent for HSR remote sensing image segmentation comprises two sub-networks: the Scale Control Agent

(SCA)  $\pi(a|s)$  and the multi-scale segmentation network  $\pi(Y|X)$ . The entire HSR image is considered as the segmentation target, and the segmentation process is modeled as a Markov Decision Making (MDP) process. The SCA selects an appropriate scale based on the *states*  $s$  of global thumbnail  $X_{\text{thumb}}$  and the position mask  $M_{\text{pos}}$ , and determines the *action*  $a$  which control the input scale of the segmentation network. The dual-branch segmentation network  $\pi(Y|X)$  takes the local image patches  $X_{\text{loc}}$  and the multi-scale context image patch  $X_{\text{context}}$  to assign categories to each pixel of the current patch. The sequence of *states*, *actions*, and *rewards* form a complete *trajectory*  $\tau$ , representing the entire mapping process of image  $X_i$ . The large scale segmentation results  $\hat{Y}_i$  are obtained and a final mapping *reward* will be given. The objective of GeoAgent is to learn an *agent*  $\pi(a|s)$  and a segmentation network  $\pi(Y|X)$  that maximize the expected cumulative *reward* of  $J(\tau)$ .

#### 3.2. Scale Control Agent Module

##### 3.2.1 Patch Scale Selection

The High Spatial Resolution (HSR) remote sensing image segmentation dataset, denoted as  $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$ , comprising  $N$  pairs of HSR remote sensing images  $X_i$  and the corresponding training labels  $Y_i$ , serves as the RL environment in this study. Each *trajectory*  $\tau$  begins with the environment’s initialization, achieved by randomly selecting an image  $X_i$  of size  $H \times W$  at its original resolution, which is subsequently divided into  $T$  fixed-size image patches. Each image patch’s segmentation defines a

timestep  $t$  in the *trajectory*.

At each timestep, the *state*  $s_t$  is represented by the global thumbnail  $X_{\text{thumb}}^{\downarrow}$  and the position mask  $M_{\text{pos}}$  of the current patch. The thumbnail  $X_{\text{thumb}}^{\downarrow}$  is a downsampled HSR image of specific size  $h \times w$  (e.g.,  $512 \times 512$ ), containing the global-scale context under coarse resolution. The position mask  $M_{\text{pos}}$  is a binary mask of size  $h \times w$ , indicating the current patch's position, with the value of the current extent set to 1 and the rest of the pixels set to 0.

As shown in Fig.2, the Scale Control Agent (SCA)  $\pi(a|s)$  is comprised of a policy network, namely, the 'actor', that determines the appropriate *action*  $a_t$  for each timestep (image patch)  $t$ . The *action* space  $a_t \in 1, 2, \dots, N$  represents the scale of the context. In addition to the local image patches  $X_{\text{loc}}$  of size  $h \times w$  with the original resolution, the context image patch  $X_{\text{context}}^{\downarrow}$  of same size  $h \times w$  with coarse resolution downsampled by a factor of  $a_t$  is acquired. The segmentation network  $\pi(Y|X)$  assigns categories to each pixel of the current patch based on  $X_{\text{loc}}$  and  $X_{\text{context}}^{\downarrow}$ . The *action*  $a_t$  receives an immediate *reward*  $r_t$  based on the segmentation result and the *environment*  $Y_i$ . Following the *action*, the *environment* transitions to the next *state*, where the sliding window moves to the next image patch. This sequence of *actions* is repeated until the entire image is observed, and the final mapping results  $\hat{Y}_i$  are obtained.

Since the difference between the *states* of the different positions lies only in position mask  $M_{\text{pos}}$ , this makes discriminability of the different *states* very small. When the *states* are directly used as input, it is difficult for SCA to recognize the weak differences and thus select similar *actions*. To improve the learning capability of the SCA, a feature indexing module is proposed to handle the subtle differences between *states*. As shown in Fig.2, the SCA is not directly trained to learn the mapping from similar *states* to different *actions*. First, the thumbnail  $X_{\text{thumb}}^{\downarrow}$  is fed to a CNN-based backbone for whole-image feature extraction. Then, the features relevant to the current patch are indexed by the position mask  $M_{\text{pos}}$ , and the irrelevant features outside the patch are masked. The features of the current patch are extracted through a  $3 \times 3$  window and transformed into feature vectors by global average pooling. Finally, an actor head with multiple fully connected layers outputs the *action* probabilities.

### 3.2.2 Segmentation Rewards and Values of Actions

After formulating the MDP of GeoAgent, the *agent*'s parameters must be optimized using DRL methods to maximize the cumulative *reward*  $J(\tau)$ . The goal of the SCA is to select the appropriate scale to obtain contextual information outside the patch and improve mapping accuracy. To penalize inappropriate *actions*, the immediate *reward* for each patch is defined as the accuracy gain of the final re-

sult based on the *action*  $a_t$  chosen by the *agent* compared to the result obtained only by the local scale, i.e.,  $a_t = 1$ :

$$\mathcal{R}_{\text{patch}}(a_t, y) = \text{Score}(y, \hat{y}(a_t)) - \text{Score}(y, \hat{y}(1)) \quad (1)$$

where  $\text{Score}(y, \hat{y}) = mIoU(y, \hat{y}) + mF_1(y, \hat{y})$ , and  $mIoU$  and  $mF_1$  are mean Intersection-Over-Union and mean  $F_1$  score, respectively. The *reward* value based only on the local image is 0, and the value is positive if the *action* leads to an accuracy improvement and negative if it results in a decline in accuracy. However, rewarding only the current single patch may cause the *agent* to overlook the potential long-term benefits of whole image mapping. Therefore, after completing the episode, the accuracy of the predicted results for the entire image is evaluated, and the final global mapping *reward* is given:

$$\mathcal{R}_{\text{map}} = T * (\text{Score}(Y_i, \hat{Y}_i) - \text{Score}(Y_i, \hat{Y}_i^1)) \quad (2)$$

where  $T$  is the number of generated patches for normalization, and  $\hat{Y}_i^1$  is the mapping result of the whole image  $X_i$  based only on the local scale.

To evaluate the expected accumulative future *reward* of a *state* or *state-action* pair, *value* function  $V^\pi(s)$  and  $Q^\pi(s, a)$  are introduced:  $V^\pi(s)$  denotes the cumulative *reward* from using policy  $\pi$  from *state*  $s$ :

$$V^\pi(s) = \mathbb{E}_{s_0=s, \tau \sim \pi} \left[ \sum_{t=0}^T \gamma^t r_t \right] \quad (3)$$

and  $Q^\pi(s, a)$  denotes the cumulative *reward* brought by performing *action*  $a$  from *state*  $s$  based on policy  $\pi$ :

$$Q^\pi(s, a) = \mathbb{E}_{s_0=s, a_0=a, \tau \sim \pi} \left[ \sum_{t=0}^T \gamma^t r_t \right] \quad (4)$$

To estimate the *value* of each *action*, in addition to the actor network to learn the mapping from current *states* to *actions*, SCA also contains a critic network that shares weights with the actor network to learn the *value* function for evaluating the current *states* and *actions*. The critic network shares a CNN backbone with the actor network to extract the features, but has a separate header to output the estimated values for later optimization of the network parameters.

### 3.3 Scale-Variable Segmentation based on Dual-Branch Network

The Dual-Branch segmentation Network, denoted as  $\pi(Y|X)$ , produces the segmentation *action* given the *state* of a local image patch  $X_{\text{loc}}$  and a multi-scale context patch  $X_{\text{context}}^{\downarrow}$ . The local patch  $X_{\text{loc}}$ , of size  $h \times w$ , is sampled from the original resolution image based on the patch position. The context patch  $X_{\text{context}}^{\downarrow}$  is obtained by expanding the local patch size by a factor of  $a_t$  centered at the current position and then downsampling it by the same factor to obtain a context patch of size  $h \times w$ .

As shown in Fig.2, the  $\pi(Y|X)$  network employs a two-branch structure sharing a common segmentation network to handle multi-scale input images. The context branch takes downsampled global thumbnails  $X_{context}^\downarrow$ , while the local branch receives local patches  $X_{loc}$  of the same size. When  $a_t = 1$ , i.e., when the *agent* believes that only local image patches are sufficient, the context branch is deactivated, and the dual-branch segmentation network is reduced to a single branch network. During the segmentation process, all layers in either branch share the same weights. The high-level feature maps  $f_l$  and  $f_c$  are then combined to generate the final segmentation mask using a branch aggregation layer.

During the feature fusion phase, the range corresponding to the local image is cropped from the context patch feature map based on its geographic coordinates. The two scales of feature maps are concatenated and fused by a convolutional layer of size  $3 \times 3$ . Finally, the segmentation probability is computed using a  $1 \times 1$  convolution and a softmax layer. To facilitate training for both branches, auxiliary headers are added to output segmentation results for local patches and context patches based on either  $f_l$  or  $f_g$  at a single scale.

### 3.4. Optimization of GeoAgent

The Advantage Actor Critic (A2C) algorithm is employed for stable learning of the *agents* in an adaptive *environment* [26]. As illustrated in Section 3.2, the SCA module consists of two learnable components: an "actor"  $\theta_A$  that learns the parameterized policy and a "critic"  $\theta_C$  that learns the *value* function to evaluate the *state-action* pair. The critic provides reinforcement signals to the actor by learning the advantage function  $A^\pi(s, a)$  [26]:

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s) \quad (5)$$

where  $V^\pi(s)$  denotes the cumulative *reward* from using policy  $\pi$  from *state*  $s$ , and  $Q^\pi(s, a)$  denotes the cumulative *reward* brought by performing *action*  $a$  from *state*  $s$  based on policy  $\pi$ .

The target advantage value  $\hat{V}_{tar}^\pi(s_t)$  can be estimated by the temporal difference (TD) algorithm and *trajectory* data (Sutton 1988), and used to train the critic network  $\theta_C$  by mean squared error (MSE) loss [26]:

$$L_{value}(\theta_C) = \frac{1}{T} \sum_{t=0}^T (\hat{V}^\pi(s_t) - V_{tar}^\pi(s_t))^2 \quad (6)$$

The actor network predict the advantage *value*  $\hat{A}^\pi(s_t, a_t)$  and optimized by the advance based on the policy loss (Mnih et al. 2016; Williams 1992):

$$L_{policy}(\theta_A) = \frac{1}{T} \sum_{t=0}^T (-\hat{A}^\pi(s_t, a_t) \log \pi \theta_A(a_t | s_t)) \quad (7)$$

The parameters of the segmentation network  $\pi(Y|X)$  can be updated through supervised learning from the loss function calculated directly from the labels. The commonly

used cross-entropy loss function is employed to optimize the segmentation network. In addition to the loss values computed from the final outputs, additional segmentation head outputs in the global and local branches are also used to compute the auxiliary losses.

When the entire GeoAgent is trained, the segmentation network  $\pi(Y|X)$  is first pre-trained by taking random *action*  $a_t$ . After completing the pre-training, the SCA  $\pi(a|s)$  is trained to learn to select the appropriate scale dynamically. Finally, joint learning is performed to optimize the segmentation network and the SCA. Since the data reading pipelines of the two training methods differ greatly,  $\theta_{scale}$  and  $\theta_{cls}$  are updated asynchronously at specific intervals to avoid the instability caused by simultaneous optimization.

## 4. Experiments

### 4.1. Experimental Setup

Three semantic segmentation datasets for HSR remote sensing images were used for experiments:

**Gaofen Image Dataset (GID)** [37]: GID is a high resolution semantic segmentation dataset based on Gaofen-2 (GF-2) satellite images. The satellite collects spectral reflectance information in the red-green-blue and near-infrared ranges to create 4-band images, with a spatial resolution of 4 meters from the multispectral camera. The dataset includes 10 training and testing images, each labeled with 15 categories and with a size of  $7200 \times 6800$ . The GID dataset is split into 50% for training and 50% for testing.

**Five-Billion-Pixels (FBP)** [38]: FBP is large-scale LULC semantic segmentation dataset. Similar to the GID dataset, the FBP dataset is built based on 4-m resolution GF-2 images annotated in a 24-category system. The FBP dataset contains 150 images with 5 billion labeled pixels, and the size of each image is also  $7200 \times 6800$ .

**Wuhan urban semantic understanding dataset (WUSU)**: We create A high-resolution LULC semantic segmentation dataset for Wuhan using GF-2 satellite imagery with finer annotations. The spatial resolution of the WUSU dataset was improved to 1-m through the fusion of panchromatic data. The categories were refined to differentiate between building instances and label them as high- or low-rise buildings, as well as to annotate excavations and structures within the city. The dataset annotates six satellite images with a total of 67,764 instances, with image heights and widths ranging from 5500 pixels to 7025 pixels. The dataset can be found in <https://github.com/AngieNikki/openWUSU>.

The GeoAgent was implemented based on PyTorch and Stable-Baseline3 [31]. The ResNet-18 is used for policy networks, and the backbones of segmentation networks are ResNet-50. The training patches were randomly sampled with the patch size of  $512 \times 512$  and batch size 16. The

Table 1: Segmentation accuracies of different models on three datasets.

Method	Backbone	WUSU		GID		FBP	
		IoU(%)	F <sub>1</sub> (%)	IoU(%)	F <sub>1</sub> (%)	IoU(%)	F <sub>1</sub> (%)
Local scale patch based methods							
-	UNet [33]	54.10	59.52	58.76	68.93	53.02	59.67
-	PSPNet [50]	60.37	69.29	59.04	69.34	59.69	64.27
-	FPN [19]	64.85	72.68	55.31	65.79	61.88	65.50
-	Deeplabv3+ [5]	63.39	69.97	64.89	68.65	57.08	59.46
-	UNet++ [54]	60.26	66.88	61.71	71.68	55.31	58.43
Global-local scale patch based methods							
GLNet [6]	FPN	51.71	59.09	58.69	68.79	21.86	24.34
MagNet [18]	FPN	51.47	63.46	66.98	55.94	45.05	49.09
RAZN [12]	Deeplabv3+	58.62	66.27	63.52	68.49	57.63	57.73
WiCoNet [11]	ResNet50	61.83	75.68	65.80	76.72	58.83	70.28
CascadePSP [7]	PSPNet	67.48	77.42	72.45	77.24	67.35	70.03
Scale-variable segmentation network GeoAgent							
GeoAgent	UNet	70.33 (+16.23)	77.11 (+17.59)	75.88 (+17.12)	75.23 (+6.30)	57.77 (+4.75)	61.37 (+1.70)
GeoAgent	UNet++	74.93 (+14.67)	79.17 (+12.29)	77.30 (+15.59)	78.62 (+6.94)	61.84 (+6.53)	60.01 (+1.58)
GeoAgent	Deeplabv3+	76.19 (+12.80)	80.56 (+10.59)	75.51 (+10.62)	76.07 (+7.42)	62.25 (+5.17)	63.21 (+3.75)
GeoAgent	PSPNet	75.19 (+14.82)	79.62 (+10.33)	75.57 (+16.53)	79.34 (+10.00)	69.67 (+9.98)	74.36 (+10.09)
GeoAgent	FPN	76.00 (+11.15)	82.46 (+9.78)	78.16 (+22.85)	77.56 (+11.77)	73.95 (+12.07)	74.11 (+8.61)

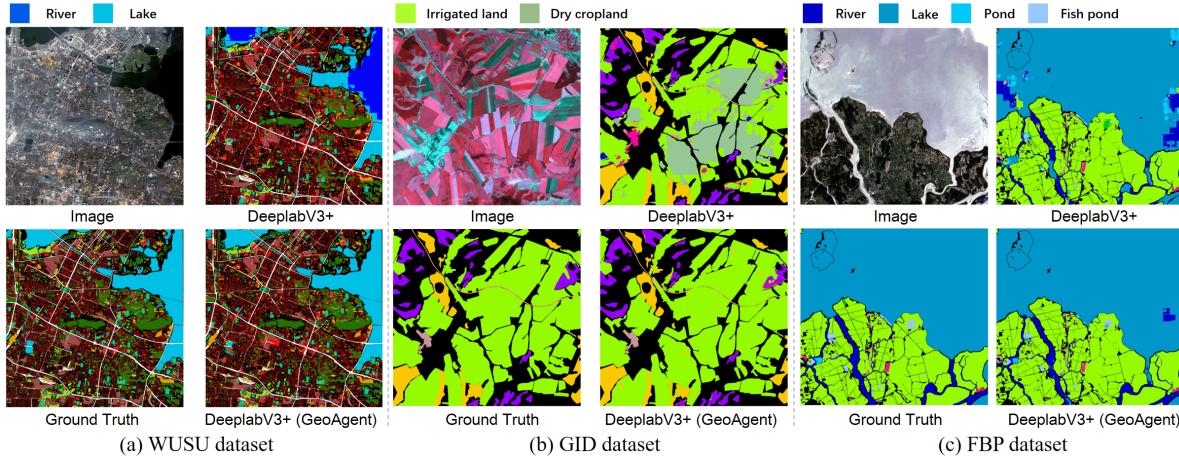


Figure 3: Segmentation results of DeeplabV3+ and modified GeoAgent models on three dataset.

stochastic gradient descent optimizer was used with the learning rate initialized to 0.001 with decay. The hyper-parameters of A2C follow the default settings of stable-baseline3, except that the batch size is adjusted to 16. The models were trained for 20,000 steps on the GID and WUSU datasets and 60,000 on the FBP dataset. The segmentation *agents* are pre-trained for 10,000 steps on the GID and WUSU datasets and 20000 on the FBP dataset. In the joint training phase, the two components perform asynchronous parameter updates at 100-step intervals. The mIoU and mean F<sub>1</sub> score were used for evaluation.

## 4.2. Comparison With Segmentation Methods

The proposed GeoAgent is a generic framework where the segmentation backbone can be an arbitrary FCN model. The GeoAgent is compared with the commonly used HSR remote sensing imagery semantic segmentation networks, including UNet [33], UNet++ [54], FPN [19], PSPNet [50], Deeplabv3+ [5]. Additionally, the global scale methods

such as GLNet [6], MagNet [18], RAZN [12], WiCoNet [11] and CascadePSP [7] are also compared.

The results listed in Table.1 show that GeoAgent significantly outperforms different patch-based based segmentation models in the challenging three HSR remote sensing image segmentation datasets. With the introduction of the GeoAgent, a high accuracy improvement can be obtained for all the models. When compared to fixed global-local scale patch-based methods such as GLNet, WiCoNet, and CascadePSP, which also employ ResNet50, SFPN, and PSPNet as backbones, GeoAgent also demonstrates accuracy advantages on all three datasets.

Examples of DeeplabV3+ on three datasets are given in Fig.3. For some artificial objects, such as roads, buildings, and residential areas, DeeplabV3+ can distinguish these categories well. However, larger natural objects, such as "rivers" and "lakes" in the WUSU and FBP datasets, have almost the same features as lakes and ponds under the finite

size  $512 \times 512$  image patch, and the "irrigated land" and "dry cropland" in the GID are also very confusing. The fixed scale image patch limits all existing models to identify different parts of the water body as different classes, leading to fragmented segmentation results. In contrast, the proposed GeoAgent can obtain more global information by capturing context outside the image patch to a suitable size based on different objects, the large-scale LULC geo-objects are correctly identified and segmentation results are consistent.

### 4.3. Ablation Experiment

In this section, the ablation study was conducted to verify the validation of the proposed GeoAgent. Different scale-selection policies of SCA are compared. All experiments are based on the Deeplabv3+, benefiting from its balanced performance and computational complexity. The GeoAgent was compared with the following strategies:

**Local only:** This baseline policy uses the traditional pipeline of FCNs with fixed  $512 \times 512$  local patches as input. **Context only:** In this policy, the datasets were down-sampled with fixed scales by 2-6 times, and the  $512 \times 512$  down-sampled context patches were directly fed into FCNs. **Single-branch:** Although the SCA module is used to select the appropriate patch scale, only one branch, a normal segmentation network, is used to process the different multi-scale image patches. This policy is similar to RAZN [12]. **Fixed scale:** The local and context branches were trained with the local image patches and context thumbnails, but the scales of the context branch are fixed to 2-6. This policy is similar to the fixed scale methods GLNet [6] and WiCoNet[11]. **Random scale:** In this policy, the local and context branches were both activated, but the scales were randomly selected. Actions were randomly performed five times to report the mean and variance. This is to verify that the DRL-based scale selection gives better results than the random scale selection. When testing, the model performs inference at all scales, and the obtained probabilities are averaged to obtain the final results.

The results for the WUSU and FBP datasets are shown in Fig.4. The segmentation accuracy of the simple down-sampled thumbnails decreases significantly as the selected scale increases, which is reasonable because the increasingly large scale downsampling directly destroys the information in the images. The global information introduced using the fixed scale performs better but negatively affects the final segmentation results in specific cases (fixed 2x scale for the FBP dataset). Moreover, the optimal fixed scales are different for different datasets. Random scale policy performs better, with a significantly higher final score than the baseline using only local information. Our proposed GeoAgent, on the other hand, dynamically decides the scale of patches and achieves the best performance.

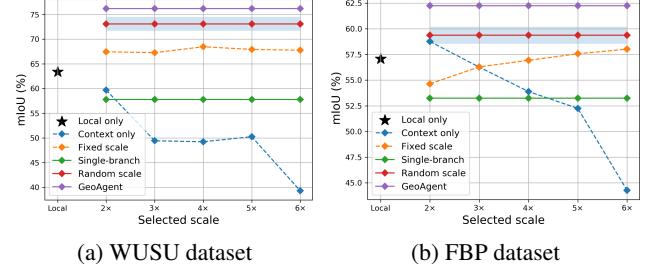


Figure 4: Ablation experiment results on different datasets.

### 4.4. Sensitivity Analysis of SCA

In this section, we investigate the impact of various reinforcement learning methods and backbones on the policy network of GeoAgent. As the FCN and DRL algorithms of the GeoAgent framework are interchangeable, we train all methods compared in Section 4.2 using different DRL algorithms for each dataset. In addition to the A2C used [26], two representative algorithms, Deep Q Network (DQN) [28] and Proximal Policy Optimization (PPO) [34], are employed for comparison. The means and standard deviations of average episode *rewards* during training with different DRL algorithms for each dataset are shown in Fig.5.

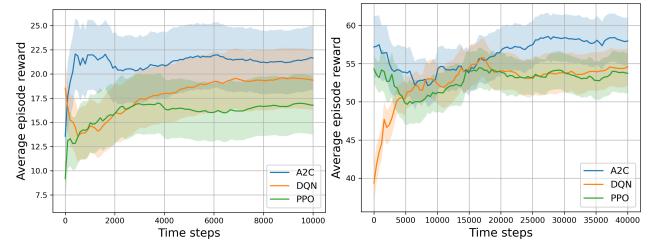


Figure 5: Average performance of various DRL methods in different datasets.

All DRL algorithms prove effective in learning scale-selection policies on all three datasets. On the all datasets, all models converge rapidly due to the relatively small size of the data volume, with A2C converging the fastest and ultimately converging to the highest value of the average episode *rewards*. A2C demonstrates an advantage in terms of convergence speed and results, although other DRL methods also perform well.

The models may prioritize short-term returns over the long-term mapping accuracy of the entire image, which is represented by  $\mathcal{R}_{map}$ . To address this, the accuracy gain of different models using GeoAgent is calculated and presented in Table.2, in addition to the average episode returns during training. GeoAgent consistently delivers significant mapping accuracy gains on all datasets. While there are some differences in performance among the DRL methods, they all result in significant accuracy gains. Despite these differences, the final mapping results are not signifi-

cantly impacted, highlighting the robustness of the GeoAgent framework to reinforcement learning algorithms.

Methods	WUSU	GID	FBP
DQN [28]	13.85±1.9	15.63±3.0	<b>7.76±3.0</b>
PPO [34]	13.51±1.8	15.77±3.0	7.76±3.0
A2C [26]	<b>13.93±1.7</b>	<b>16.54±3.1</b>	7.70±2.9

Table 2: Accuracy improvement of various FCNs for different DRL methods in the three datasets.

#### 4.5. Visualization Analysis

This section analyzes the *actions* of SCA, which are recorded by sliding a window through an image and completing an episode. The resulting *action* maps of GeoAgent models based on A2C and Deeplabv3+ are presented in Fig. 6 for three datasets. To evaluate the effectiveness of the proposed feature indexing method, the *action* maps of the policy network without this module are compared. In this case, thumbnails and position mask *states* are concatenated to form a multi-channel image, which is then fed to the CNN for feature extraction and output *action* probabilities.

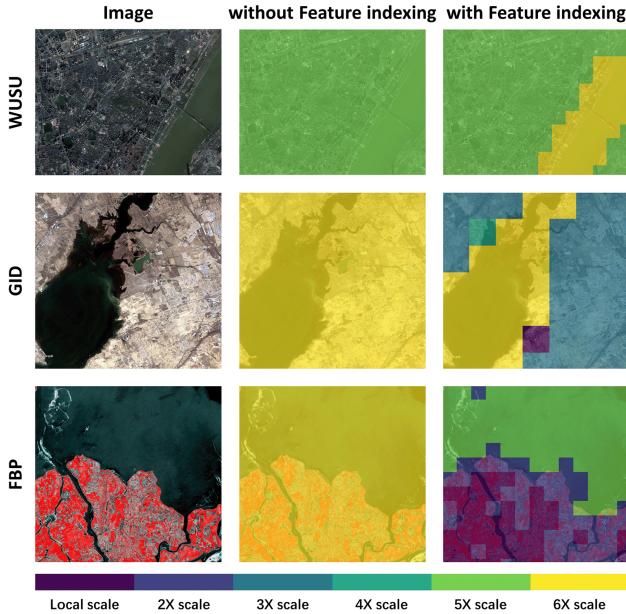


Figure 6: Action map visualization of GeoAgent on different datasets.

The results indicate that the GeoAgent can adaptively take different context scale for different *states*. For large-scale objects, such as rivers and lakes, larger scale chosen to obtain more background information due to the lack of long-range spatial relationships at smaller observation scales. The SCA mainly take larger magnification zoom for natural objects with large scales, while smaller scales or even no zooming are chosen for artificial features that can be easily identified.

Without the feature indexing module, the policy network finds it difficult to locate the position of the current *state* on the whole thumbnail, as the position mask is fed in only as a channel of observation, and the thumbnails of different *states* are identical. Consequently, the policy network without feature indexing takes the same *action* for all *states* on the whole image. In the two examples of large-scale objects on the GID dataset and the FBP dataset, the conservative *action* of taking the largest scale on the whole image is chosen by the *agent* to guarantee a higher average expected return. This outcome highlights the effectiveness of the proposed feature indexing method.

In addition, to evaluate the generalization of GeoAgent, we trained an Agent with RGB input on the WUSU dataset and performed inference on unseen WorldView-based datasets, namely LovaDA [42], xView[20], and DeepGlobe[8]. The action maps in Fig.7 demonstrate GeoAgent’s ability to learn action strategies that generalize well across different datasets.

GeoAgent consistently chooses larger scales for big objects to understand their context, while it selects localized scales for smaller objects, showing its ability to focus on detail. This pattern is observed regardless of the dataset’s origin, proving its effective generalization skills.

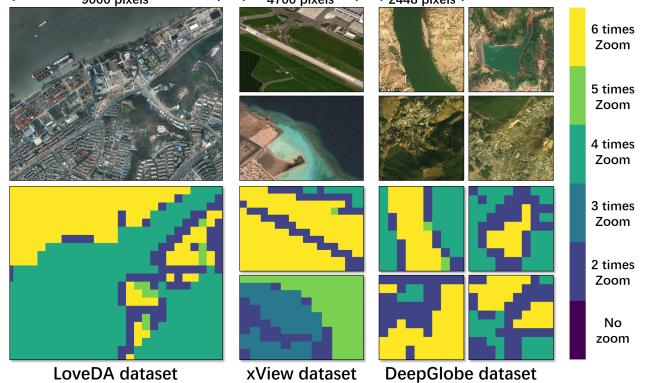


Figure 7: Action visualization of the GeoAgent on unseen datasets.

### 5. Conclusion

In this paper, we propose a deep reinforcement learning-based scale-adaptive semantic segmentation network called GeoAgent, breaking the limitation sliding window size of that conventional segmentation networks. Experiments on three datasets show that our proposed GeoAgent achieves state-of-the-art results on the semantic segmentation task of high-resolution satellite images.

### 6. Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No. 42325105, 42071350, and LIESMARS Special Research Funding.

## References

- [1] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 4(11):e21, 2019.
- [2] Kumar Ayush, Burak Uzkent, Kumar Tanmay3 Marshall Burke2 David Lobell, and Stefano Ermon. Efficient poverty mapping from high resolution remote sensing images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12–20, 2021.
- [3] Benjamin Bechtel, Paul J Alexander, Jürgen Böhner, Jason Ching, Olaf Conrad, Johannes Feddema, Gerald Mills, Linda See, and Iain Stewart. Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information*, 4(1):199–219, 2015.
- [4] Jon Atli Benediktsson, Jocelyn Chanussot, and Wooil M. Moon. [point of view]. *Proceedings of the IEEE*, 100(6):1907–1910, 2012.
- [5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [6] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8924–8933, 2019.
- [7] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8890–8899, 2020.
- [8] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018.
- [9] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020.
- [10] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7778–7796, 2021.
- [11] Lei Ding, Dong Lin, Shaofu Lin, Jing Zhang, Xiaojie Cui, Yuebin Wang, Hao Tang, and Lorenzo Bruzzone. Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.
- [12] Nanqing Dong, Michael Kampffmeyer, Xiaodan Liang, Zeya Wang, Wei Dai, and Eric Xing. Reinforced auto-zoom net: towards accurate and fast breast cancer segmentation in whole-slide images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 317–325. Springer, 2018.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [14] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatian, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.
- [15] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, 2021.
- [16] Bo Huang, Bei Zhao, and Yimeng Song. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214:73–86, 2018.
- [17] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.
- [18] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16755–16764, 2021.
- [19] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019.
- [20] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Doolley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.
- [21] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4217–4226, June 2021.
- [22] Zuoyue Li, Jan Dirk Wegner, and Aurelien Lucchi. Topological map extraction from overhead images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

- [23] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.
- [24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [26] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Tim Harley, and Timothy P. Lillicrap. Asynchronous methods for deep reinforcement learning. In MF Balcan and KQ Weinberger, editors, *INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 48*, volume 48 of *Proceedings of Machine Learning Research*, 2016. 33rd International Conference on Machine Learning, New York, NY, JUN 20-22, 2016.
- [27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [29] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [30] Nicholas J Murray, Thomas A Worthington, Pete Bunting, Stephanie Duce, Valerie Hagger, Catherine E Lovelock, Richard Lucas, Megan I Saunders, Marcus Sheaves, Mark Spalding, et al. High-resolution mapping of losses and gains of earth’s tidal wetlands. *Science*, 376(6594):744–749, 2022.
- [31] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.
- [32] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 51–66, 2018.
- [33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [35] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [36] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [37] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020.
- [38] Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *arXiv preprint arXiv:2209.00727*, 2022.
- [39] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12345–12354, 2020.
- [40] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.
- [41] Junjue Wang, Ailong Ma, Yanfei Zhong, Zhusu Zheng, and Liangpei Zhang. Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery. *Remote Sensing of Environment*, 277:113058, 2022.
- [42] Junjue Wang, Zhusu Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.
- [43] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Endre Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021.
- [44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
- [45] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. Isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.
- [46] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8817–8826, 2018.
- [47] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beßongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

- [48] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.
- [49] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.
- [50] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [51] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [52] Yanfei Zhong, Yu Su, Siqi Wu, Zhendong Zheng, Ji Zhao, Ailong Ma, Qiqi Zhu, Richen Ye, Xiaoman Li, Petri Pellikka, et al. Open-source data-driven urban land-use mapping integrating point-line-polygon semantic objects: A case study of chinese cities. *Remote Sensing of Environment*, 247:111838, 2020.
- [53] Lichen Zhou, Chuang Zhang, and Ming Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [54] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.