frontiers | Frontiers in Environmental Science

TYPE Original Research
PUBLISHED 17 November 2022
DOI 10.3389/fenvs.2022.1010630

OPEN ACCESS

EDITED BY
Penghai Wu,
Anhui University, China

REVIEWED BY
Junli Li,
Anhui Agricultural University, China
Guojie Wang,
Nanjing University of Information Science and Technology, China

*CORRESPONDENCE
Huihui Feng,
hhfeng@csu.edu.cn

SPECIALTY SECTION
This article was submitted to Environmental Informatics and Remote Sensing, a section of the journal Frontiers in Environmental Science

RECEIVED 03 August 2022
ACCEPTED 07 November 2022
PUBLISHED 17 November 2022

CITATION
Lu W, Qi J and Feng H (2022), Urban functional zone classification based on self-supervised learning: A case study in Beijing, China.
Front. Environ. Sci. 10:1010630.
doi: 10.3389/fenvs.2022.1010630

Urban functional zone classification based on self-supervised learning: A case study in Beijing, China

Weipeng Lu, Ji Qi and Huihui Feng*

School of Geosciences and Info-Physics, Central South University, Changsha, China

Urban functional zones (UFZs) are the fundamental units for urban management and operation. The advance in earth observation and deep learning technology provides chances for automatically and intelligently classifying UFZs *via* remote sensing images. However, current methods based on deep learning require numerous high-quality annotations to train a well-performed model, which is time-consuming. Thus, how to train a reliable model using a few annotated data is a problem in UFZ classification. Self-supervised learning (SSL) can optimize models using numerous unannotated data. In this paper, we introduce SSL into UFZ classification to use the instance discrimination pretext task for guiding a model to learn useful features from over 50,000 unannotated remote sensing images and fine tune the model using 700 to 7,000 annotated data. The validation experiment in Beijing, China reveals that 1) using a few annotated data, SSL can achieve a kappa coefficient and an overall accuracy 2.1–11.8% and 2.0–10.0% higher than that of supervised learning (SL), and 2) can also gain results comparable to that got by the SL paradigm using two times annotated data for training. The less the data used for finetuning the more obvious the advantage of SSL to SL. Besides, the comparison experiment between the model pretrained on the research region and that pretrained on the benchmark reveals that the objects with displacement and incompleteness are more difficult for models to classify accurately.

KEYWORDS

self-supervised learning, urban functional zone, remote sensing, deep learning, image classification

1 Introduction

Urban functional zones (UFZs), including commercial zones, industrial zones, and residential zones, have specific social activities. The spatial distribution of UFZs describes the city structure and reveals the land demand, playing an important role in urban management (Zhang et al., 2017; Chen et al., 2018). Nowadays, geographic big data like points of interest and geo-tagged photos become available, which were used to analyze UFZ spatial patterns. For example, Yin et al. (2021b) used the density of points of interest

to determine the type of parcels and map out the UFZ. Kang et al. (2021) used photos from Flickr to investigate the landscapes to guide the tourisms industry. However, these data were uploaded by users, so their quality are uncontrollable (Yin et al., 2021a). The advance in earth observation provides high spatiotemporal-resolution remote sensing imagery (RSI), which is widely used for UFZ classification research (Bao et al., 2020; Cao et al., 2020; Liu et al., 2021).

Traditional RSI interpretation relies on handcrafted features (Dai and Yang, 2010; Zhu et al., 2014; Castelluccio et al., 2015), in which radiometric features, texture features, and shape features were used for image classification and retrieval (Luo et al., 2013). Zhang et al. (2018) proposed a hierarchical bottom-up and up-bottom feedback model to improve the classification accuracy of UFZs by handcrafted features like gray-level co-occurrence matrix (GLCM). Du et al. (2019) used window independent context (WIC) feature to extract spatial units of UFZs from very-high-resolution RSI. However, generating a well-designed handcrafted feature requires expert experience and has low robustness, which cannot provide satisfying results in complex RSI interpretation like UFZ classification (Cheng et al., 2017).

Recently, with the development of deep learning technology, the methods based on high-level visual features, like convolutional neural networks (CNNs), are employed in intelligent and automatic feature extraction (Ioffe and Szegedy, 2015; He et al., 2016; Szegedy et al., 2016). More and more UFZ researchers have adapted CNNs for representation and classification (Liu et al., 2017; Cheng et al., 2018; Wang et al., 2018). For UFZ classification, CNNs have become an essential part in recent 5 years (Bao et al., 2020; Liu et al., 2020; Xu et al., 2020; Zhou et al., 2020; Du et al., 2021; Lu et al., 2022). Zhou et al. (2020) proposed super-object based CNNs to classify UFZ in RSI. They used the AlexNet (Krizhevsky et al., 2012), a typical CNN model, to determine the class of a clipped RSI. Du et al. (2021) designed a multi-scale semantic segmentation network combining an object-level conditional random field to map UFZ at the object level.

Generally, the training of a CNN follows the supervised learning (SL) paradigm, which fits the parameters using numerous annotated training data. Under the SL paradigm, training a stable model requires a large number of high-quality samples (Ma et al., 2017). Large-scale image classification datasets, such as ImageNet (Krizhevsky et al., 2012) and **P**attern **A**nalysis, **S**tatical modeling and **C**omput**A**tional **L**earning **V**isual **O**bject **C**lasses (PASCAL VOC) challenge dataset (Everingham et al., 2010) have promoted the development of SL in computer vision. However, when SL is applied to the field like remote sensing and medical image, this training paradigm often has insufficient training samples. Annotating an RSI dataset needs professional knowledge and tedious work, so annotating an RSI dataset as large as ImageNet is costly. Therefore, it is difficult to train a good-performance UFZ classification model with existing datasets under the SL paradigm.

Transfer learning (TL) pretrains a model on large-scale datasets *via* SL and then finetunes parts of model parameters by target tasks, like UFZ classification. It can reduce the annotation requirement (Wang et al., 2020; Yang et al., 2020). TL assumes that the model can learn a general representation from large amounts of datasets. And the representation can be transferred into the remote sensing domain by a few annotated

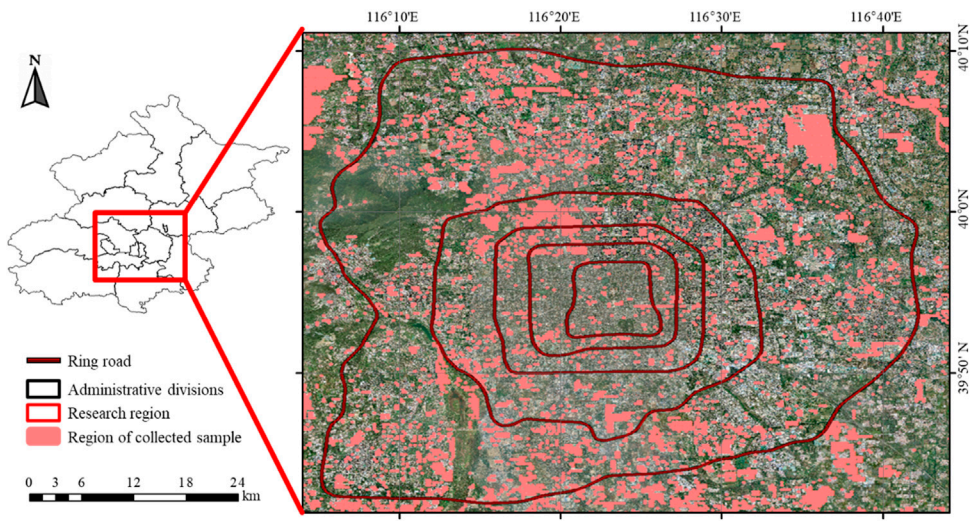

FIGURE 1
The study area and its location in Beijing, China.

TABLE 1 The classification system and number of patches for each UFZ type of the collected dataset.

Category	Definition	Number of patches
Commercial	Financial center, retail center, shopping mall, office building	314
Residential	Residence, urban shantytown, and rural settlement	1226
Institutional	educational, medical, cultural, administrative office, and public services	763
Industrial	factories, warehouse	275
Transportation	Railway, highway, port and its surrounding water, bus station, railway station, airport, gasoline station	875
Open Space	urban park, botanic garden, and other urban grasslands	531
Construction	vacant land, bare land, and land under construction	1310
Forest	non-urban development land with dense trees	333
Agricultural	vegetable field, cropland, orchard, and other agricultural lands	1105
Water	natural and artificial waterbody	572
Σ	7304	

data. But TL requires that the data used for pretraining and finetuning should have the same number of channels. Natural images have the three channels of red band, green band, and blue band (RGB bands), but different RSIs have different numbers of channels. For example, multispectral images and hyperspectral images have more than three channels and panchromatic imagery has only one channel. The difference in channel numbers causes difficulty in finetuning the RGB-pretrained model on RIS. In addition, the RGB-band RSIs have quite different visual characteristics from natural images, due to the different imaging mechanisms, such as angle and distance. Therefore, it is a problem to train a model *via* massive unannotated RSIs.

In the past few years, self-supervised learning (SSL) has become popular in model pretraining and gains results comparable to those got by previous learning paradigms in computer vision tasks such as image classification, semantic segmentation, and object detection (Doersch and Zisserman, 2017; Similarities, 2021; Tao et al., 2021; Li et al., 2022). SSL trains models to learn useful knowledge *via* pretext task, whose annotation is obtained directly from the training data. Thus, SSL has the advantage that its pretrain period is label-free. Recently, SSL researches on RSI have made great progress, but most of them only used public benchmark for experiments (Yu et al., 2020; Zhao et al., 2020; Stojnic and Risojevic, 2021). For example, Tao et al. (2022) investigated the potential of SSL on RSI interpretation by three open RSI datasets: **EuroSAT**, which is the Land Use and Land Cover Classification with Sentinel-2 (Helber et al., 2019), **A**erial **I**mage **D**ataset **(AID)** (Xia et al., 2017), and **NWPU-RESISC45**, which is the **RE**mote **S**ensing **I**mage **S**cene **C**lassification (RESISC) dataset created by **N**orthwestern **P**olytechnical **U**niversity (NWPU) with **45** classes (Cheng et al., 2017). No relevant studies have been carried out in the practical application. The RSIs selected in open RSI datasets and those used for practical applications are different (Cheng et al., 2017, 2020; Hong, 2021).

- First, in the images selected for open RSI datasets, the objects of interest are always at the center. In the images used for practical application, the location of key objects is random, so the image is difficult to be cropped with the target at the center. The displacement of objects causes sample misclassification by the model pre-trained on benchmark.
- Second, the image size of a benchmark is always fixed, but the scales of objects in benchmarks are different (e.g., factory and airport). Thus, the benchmark spatial resolution changes to make sure the key object is contained in the image completely. However, in practice, the spatial resolution and the image size are always fixed, so some large-scale object might be cropped into several patches, which is difficult for the model to classify accurately.

Therefore, this paper intends to introduce the SSL into the UFZ classification of the region inside the Sixth Ring Road of Beijing, China, and to investigate the different performance of SSL in open RSI dataset and the practical application. Specifically, we pretrain the model on unannotated RSI of downtown Beijing *via* SSL, and then collect a small-scale UFZ classification dataset to fine-tune the model. Notably, in order to be like the practical application, all samples used in the experiments are randomly cropped with fixed resolution and size. The experiment result shows that SSL has advantages over SL in terms of sample demand and final classification accuracy.

2 Materials and methods

2.1 Study area and data

This study takes Beijing, China as the research region (shown in Figure 1). It has a spatial coverage of 3300 km^2 (longitude

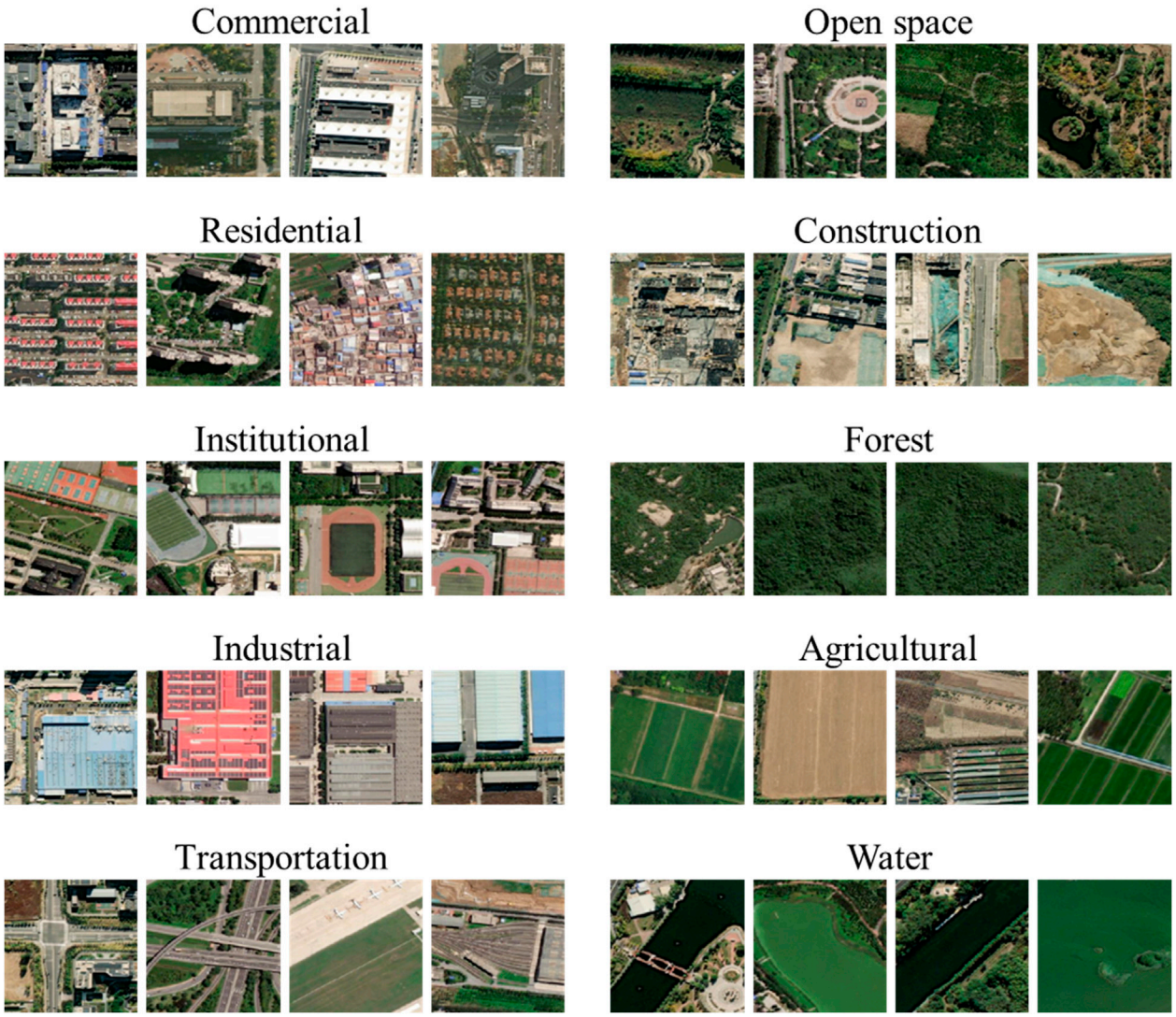

FIGURE 2
Samples of the 10 urban function zone (UFZ) types.

116°04′-116°44′E and latitude 39°40′-40°11′N) and a population of 21,000,000. This region contains a variety of urban landscapes, which can effectively denote old/new city areas and urban/suburbs.

The RSI used in this paper is downloaded from Bing Virtual with a size of 53248×69632 in the WGS-84 framework. For SSL, the entire image is meshed into 56,576 patches with a size of 256. Considering current researches (Zhang et al., 2020; Liu et al., 2021; Lu et al., 2022) and the "Code for classification of urban and rural land use and planning standards of development land (GB50137)" issued by the Ministry of Housing and Urban-Rural Development of the People's Republic of China, we divide the UFZs into 10 kinds. For model finetuning, we annotate a few patches manually. The classification system and the number of patches for each UFZ type are shown in Table 1, and parts of annotated patches are shown in Figure 2.

2.2 Paradigm of SL and SSL

The supervised learning (SL) is a model training paradigm that has been widely used in big data analysis. Given dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{k}$ and model $\mathcal{F}: x_i \rightarrow \hat{y}_i$ with random initialization parameters, SL is to optimize $\mathcal{F}$ to minimize the error between y_i and $\hat{y}_i$.

The SSL is to initialize a model's parameters using pretext tasks, such as image reconstruction, rotation prediction, and

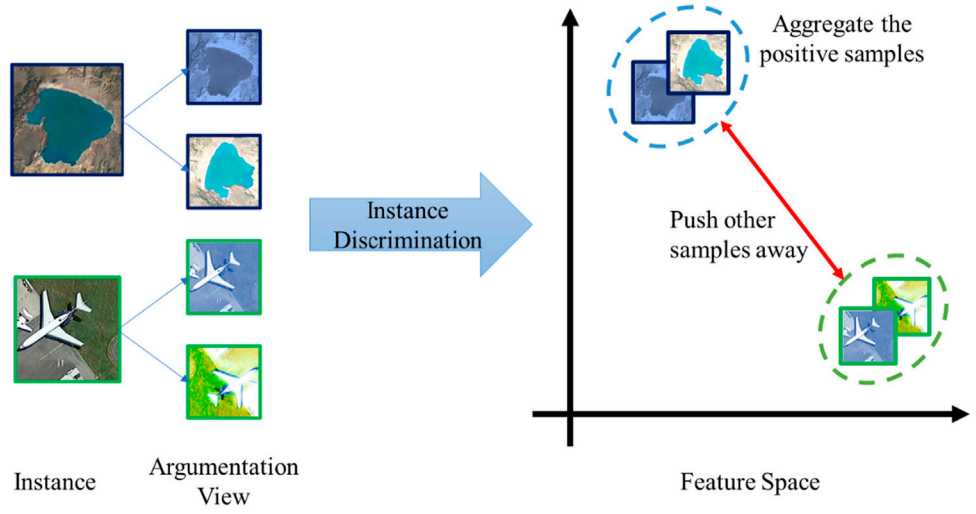

FIGURE 3
Understanding the instance discrimination task from the perspective of feature space.

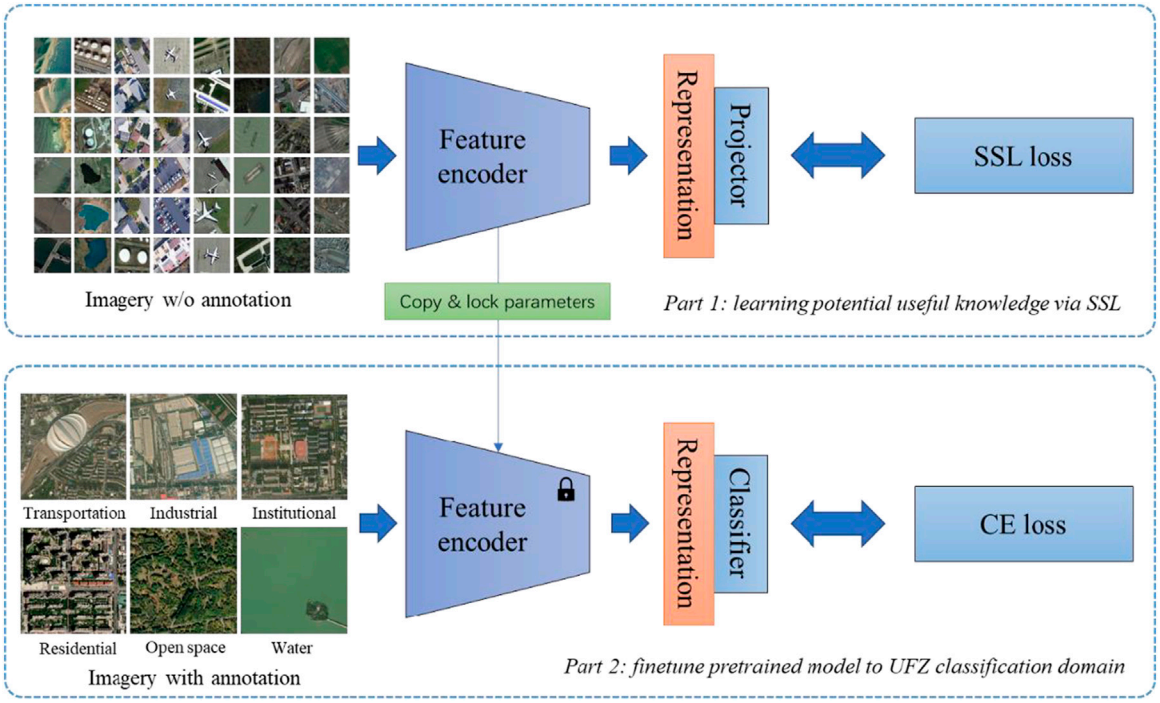

FIGURE 4
Flowchart of UFZ classification by SSL.

instance discrimination (Tao et al., 2020). By solving the pretext tasks, the model can learn the useful features from unannotated samples. Here, we introduce the instance discrimination task that will be used in our research.

Given an image (instance) $\tilde{\mathbf{x}}$ and its two argumentation views $\mathbf{x}_i$ and $\mathbf{x}_j$, instance discrimination is to distinguish the positive sample of $\mathbf{x}_i$ from a set of samples $\{\mathbf{x}_k\}$. $\mathbf{x}_i$ and $\mathbf{x}_j$ are positive samples of each other. From the perspective of feature space

TABLE 2 Evaluation of the models using SSL initialization and random initialization using different percentage of finetune samples. Com: commercial, Res: residential, Ins: institutional, Ind: industrial, Tra: transportation, OS: open space, Con: construction, For: forest, Agr: agricultural.

Initialization	Sample (%)	UA (%)/PA (%)/F1										Kappa	OA (%)
		Com	Res	Ins	Ind	Tra	OS	Con	For	Agr	Water		
SSL on the Research Region	100	57.9	85.3	67.5	74.1	83.6	81.6	77.9	98.5	90.3	95.3	0.796	82.2
		34.9	87.8	71.9	72.7	82.2	84.0	83.2	95.6	87.9	94.4		
		0.436	0.865	0.696	0.734	0.829	0.828	0.804	0.970	0.891	0.949		
	80	51.4	83.9	65.9	75.5	85.2	80.8	77.7	98.5	89.8	92.3	0.790	81.7
		30.2	86.9	71.9	72.7	82.3	79.2	82.4	97.0	87.8	94.7		
		0.380	0.854	0.688	0.741	0.837	0.800	0.800	0.977	0.888	0.935		
	40	40.9	79.1	60.6	70.8	82.5	79.8	73.3	98.5	89.4	91.2	0.753	78.5
		14.3	88.2	71.2	61.8	75.4	78.3	80.5	95.5	84.2	90.4		
		0.212	0.834	0.655	0.660	0.788	0.790	0.767	0.970	0.867	0.907		
	20	40.0	77.1	58.3	74.1	77.8	82.5	67.0	98.5	86.8	91.9	0.727	76.4
		3.2	86.5	64.1	36.4	70.3	75.5	85.9	95.5	86.0	89.5		
		0.059	0.815	0.611	0.488	0.739	0.788	0.753	0.970	0.864	0.907		
	10	0	73.3	53.6	0	79.3	92.2	52.6	98.4	75.4	90.7	0.642	69.3
		0	85.3	48.4	0	54.9	55.7	87.4	91.0	84.6	86.0		
		-	0.789	0.509	-	0.649	0.694	0.657	0.946	0.797	0.883		
Random initialization	100	35.0	83.4	67.3	70.6	81.0	71.8	85.3	97.0	89.6	95.3	0.779	80.6
		33.3	82.0	73.9	65.5	82.9	79.2	84.4	97.0	86.0	89.5		
		0.341	0.827	0.704	0.679	0.819	0.753	0.848	0.970	0.878	0.923		
	80	29.8	81.6	64.8	81.8	82.4	67.5	82.3	95.6	91.4	95.2	0.773	80.2
		22.2	83.3	67.3	65.5	82.9	76.4	88.9	97.0	86.9	86.8		
		0.255	0.824	0.660	0.727	0.826	0.717	0.855	0.963	0.891	0.908		
	40	34.8	74.7	57.2	57.1	73.3	69.6	76.7	96.8	85.8	91.6	0.709	74.7
		25.4	84.5	59.5	36.4	75.4	75.5	80.5	91.0	79.2	86.0		
		0.294	0.793	0.583	0.444	0.744	0.724	0.786	0.938	0.824	0.887		
	20	33.7	69.6	52.7	71.4	71.3	60.9	69.7	93.8	75.6	82.0	0.636	68.2
		47.6	81.2	50.3	36.4	68.0	66.0	63.4	89.6	74.2	79.8		
		0.395	0.750	0.515	0.482	0.696	0.633	0.664	0.916	0.749	0.809		
	10	31.0	65.6	44.7	50.0	72.3	55.6	58.3	76.6	72.3	85.9	0.574	63.0
		34.9	80.8	44.4	18.2	53.7	47.2	66.8	88.1	71.9	74.6		
		0.328	0.724	0.446	0.267	0.616	0.510	0.623	0.819	0.721	0.798		

(Figure 3), the goal of instance discrimination is to aggregate the positive samples and push apart them from other samples (negative samples). A similarity loss function is designed to complete the task Eq. 1.

$$l_1(i,j) = -log\frac{exp\left(cos < \boldsymbol{z}_i, \boldsymbol{z}_j > / \tau\right)}{\sum_{k=1}^{2N} \mathbb{I}_{k\neq i} exp\left(cos < \boldsymbol{z}_i, \boldsymbol{z}_j > / \tau\right)} \quad (1)$$

where $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ are the argumentation views. $cos < \boldsymbol{u}, \boldsymbol{v} > = \frac{\boldsymbol{uv}}{|\boldsymbol{u}||\boldsymbol{v}|}$. $\mathbb{I}_{bool}$ is the indicator function, and its value is 1 only if bool is true, 0 otherwise. τ is the temperature parameter.

2.3 Implementation of SSL on UFZ classification

As shown in Figure 4, the implementation of SSL on UFZ classification includes two steps: 1) learning useful knowledge *via* SSL, and 2) finetuning the pre-trained model to the UFZ classification domain *via* SL. In the first step, the model will be trained on large-scale unannotated RSIs to learn useful knowledge. In the second step, the model will be finetuned on a small-scale UFZ classification dataset with annotation to obtain a UFZ classification model.

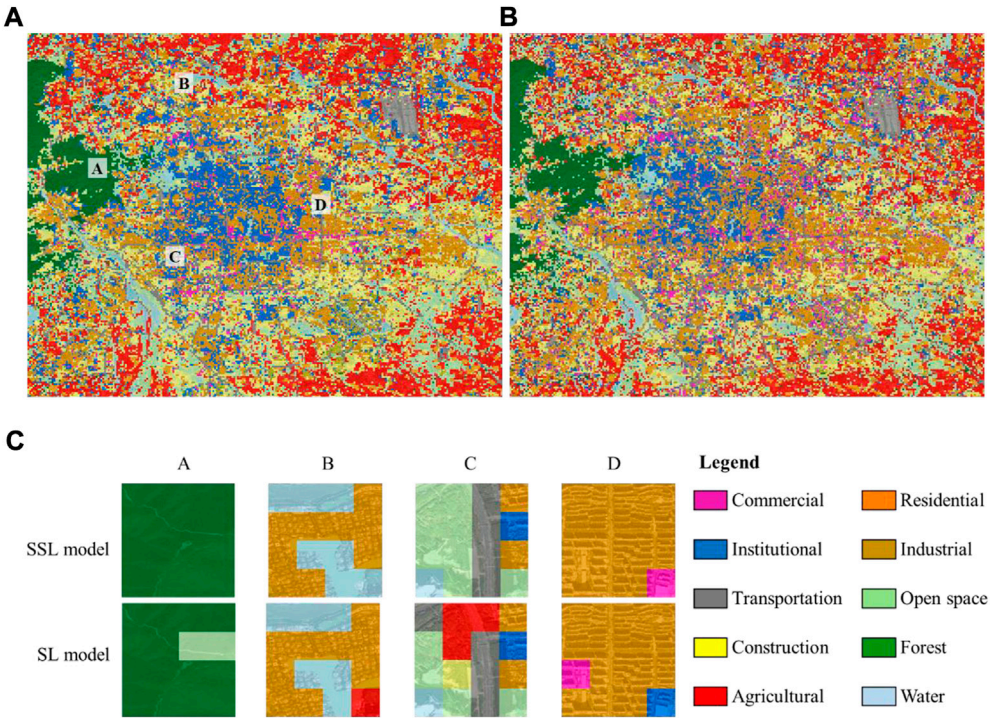

FIGURE 5
UFZ classification using 100% training samples. **(A)** UFZ map predicted by the SSL model; **(B)** UFZ map predicted by the SL model; **(C)** comparison result between the SSL model and the SL model.

2.3.1 Learning potential useful knowledge *via* SSL

In this study, we use instance discrimination as the pretext task, as it can guide the model to learn the invariance of an image and the difference between two images (Chen et al., 2020).

We design a CNN that contains a visual feature encoder $f(\cdot|\boldsymbol{\theta}_f)$ and a feature projector $g(\cdot|\boldsymbol{\theta}_g)$ to represent argumentation views and complete the instance discrimination task. The SSL training has three steps:

1) Generation of positive samples: Randomly select a few unannotated data $\{\tilde{\mathbf{x}}_k\}_{k=1}^{N}$ from a large-scale dataset $\tilde{\mathbf{X}}$, and argument them by two random argumentation rules t_1 and t_2 (e.g., rotation, flip, random mask, dithering). By doing so, a set of argumentation views $\{\mathbf{x}_k\}_{k=1}^{2N}$ are generated, in which $\mathbf{x}_{2k-1}$ and $\mathbf{x}_{2k}$ are a pair of positive samples.
2) Representation of argumentation views: represent the argumentation views in $\{\mathbf{x}_k\}_{k=1}^{2N}$ by $f(\cdot|\boldsymbol{\theta}_f)$ to get the visual representation $\{\boldsymbol{h}_k\}_{k=1}^{2N}$ and project the representation by $g(\cdot|\boldsymbol{\theta}_g)$. In this way, all argumentation views are projected as $\{\boldsymbol{z}_k\}_{k=1}^{2N}$ in the instance discriminative space.
3) Discrimination of instance: optimize $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$ by minimizing the similarity loss.

2.3.2 Finetune the pre-trained model to the UFZ classification domain

Finetuning the pre-trained model *via* SL is to use a small-scale annotated dataset to adjust some parameters of the pre-trained model. In this study, we use the collected UFZ classification dataset to finetune the pre-trained model to the UFZ classification domain through the following two steps:

1) Extracting useful features: randomly sample a mini batch of data $\{(\tilde{\mathbf{x}}_k, \boldsymbol{y}_k)\}_{k=1}^{n}$ from the annotated dataset $(\tilde{\mathbf{X}}, \tilde{\boldsymbol{Y}})$, and extract features by pretrained feature encoder $f(\cdot|\boldsymbol{\theta}_f)$ to get the feature representation $\{\boldsymbol{h}_k\}_{k=1}^{n}$.
2) Finetuning the model by SL: randomly initialize a classifier $\varphi(\cdot|\boldsymbol{\theta}_\varphi)$ and classify $\{\boldsymbol{h}_k\}_{k=1}^{n}$ to predict the class distribution probability $\{\boldsymbol{p}_k\}_{k=1}^{n}$ and optimize φ by minimizing the cross-entropy (CE) loss between $\boldsymbol{p}_k$ and $\boldsymbol{y}_k$. $y_{k,i} = \mathbb{I}_{\mathbf{x}_k \in Class_i}$, $p_{k,i} = P(\mathbf{x}_k \in Class_i)$, and cls is the total class number. In this study, cls is 10.

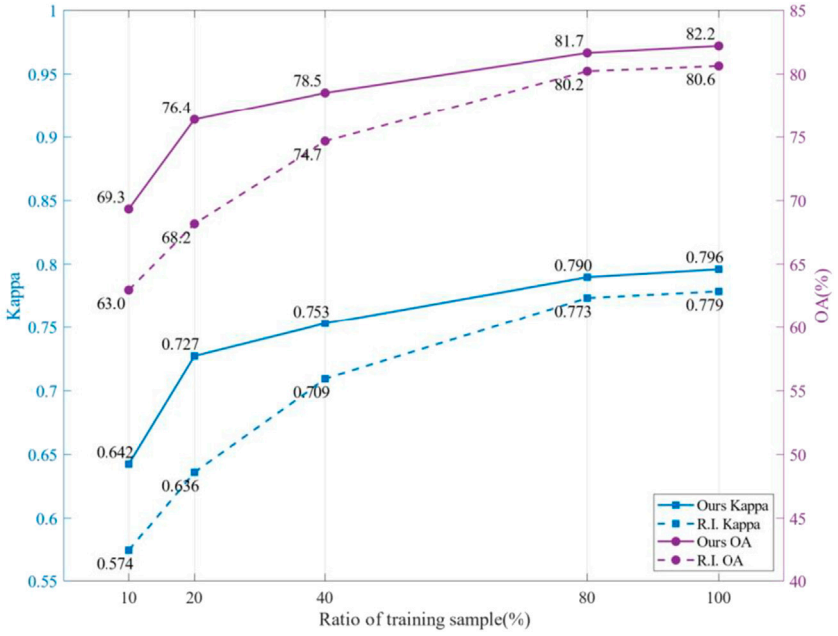

FIGURE 6
Results obtained using different ratio of training samples. R.I., random initialization.

TABLE 3 Location quotient based on the ring road.

Category	Inside 2nd	2nd-3rd	3rd-4th	4th-5th	5th-6th
Commercial	2.01	2.42	2.51	1.67	0.59
Residential	1.81	1.41	1.69	1.15	0.85
Institutional	2.82	3.04	2.00	1.27	0.66
Industrial	0.17	0.18	0.59	0.93	1.13
Transportation	0.70	1.01	1.22	1.17	0.95
Open Space	0.27	0.18	0.30	1.13	1.11
Construction	0.20	0.44	0.65	1.11	1.07
Forest	0.00	0.00	0.00	0.01	1.41
Agricultural	0.01	0.00	0.03	0.23	1.36
Water	0.85	0.26	0.19	0.67	1.20
$\mu \pm \sigma$	0.88 ± 0.94	0.89 ± 1.02	0.92 ± 0.84	0.93 ± 0.47	1.03 ± 0.26

$$l_2(\boldsymbol{p}_k, \boldsymbol{y}_k) = -\sum_{i=1}^{cls} y_{k,i} log\left(p_{k,i}\right) \tag{2}$$

2.3.3 Implementation details

In the experiment, we use ResNet50 (He et al., 2016) as the backbone of the visual encoder $f(\cdot|\boldsymbol{\theta}_f)$, and take two stacked fully connected (FC) layer with **RE**ctified **L**inear **U**nit (*ReLU*) activating function as the projector $g(\cdot|\boldsymbol{\theta}_g)$. For an image $\tilde{\mathbf{x}}_k$ (or its argumentation view $\mathbf{x}_k$), the model firstly extracts its visual feature by feature extractor $f: \tilde{\mathbf{x}}_k \rightarrow \boldsymbol{h}_k \in \mathbb{R}^{2048}$, and then projects $\boldsymbol{h}_k$ to the instance discriminative space by projector g (Eq. 3), in which $\mathbf{W}_1 \in \mathbb{R}^{1024\times2048}$ and $\mathbf{W}_2 \in \mathbb{R}^{128\times1024}$ are the learnable weight in FC layers, and $\delta(a) = max(0, a)$ denotes the *ReLU* function.

$$\boldsymbol{z}_k = g\left(h_k|\boldsymbol{\theta}_g\right) = \mathbf{W}_2\delta(\mathbf{W}_1\boldsymbol{h}_k) \tag{3}$$

For model finetuning, we take a classifier with an FC layer. Mathematically, the classification process can be expressed by (Eq. 4).

TABLE 4 Location quotient based on administrative district.

Category	Xicheng	Dongcheng	Haidian	Chaoyang	Shijing	Fengtai
Commercial	1.50	1.52	0.74	1.29	0.35	1.10
Residential	1.73	1.67	0.81	1.12	1.25	1.01
Institutional	2.69	2.58	1.39	0.77	0.69	0.90
Industrial	0.11	0.32	0.41	1.24	1.09	1.44
Transportation	0.82	0.79	0.67	1.17	0.84	1.25
Open Space	0.11	0.31	1.16	0.70	0.52	1.38
Construction	0.16	0.32	0.56	1.42	0.89	1.01
Forest	0.00	0.00	1.83	0.00	4.03	0.53
Agricultural	0.00	0.02	1.35	1.06	0.07	0.68
Water	1.23	0.20	0.89	1.37	0.58	0.70
$\mu \pm \sigma$	0.83 ± 0.88	0.77 ± 0.82	0.98 ± 0.42	1.01 ± 0.41	1.03 ± 1.05	1 ± 0.29

TABLE 5 Quantitative result of the models pretrained on the research region and AID.

Initialization	UA (%)/PA (%)/F1										Kappa	OA
	Com	Res	Ins	Ind	Tra	OS	Con	For	Agr	Water		
SSL on the Research Region	57.9	85.3	67.5	74.1	83.6	81.6	77.9	98.5	90.3	95.3	0.796	82.2 (%)
	34.9	87.8	71.9	72.7	82.2	84.0	83.2	95.6	87.9	94.4		
	0.436	0.865	0.696	0.734	0.829	0.828	0.804	0.970	0.891	0.949		
SSL on AID	46.2	69.0	64.9	68.4	79.7	77.3	63.0	90.0	79.9	90.3	0.779	80.6 (%)
	19.1	87.4	55.6	47.3	67.4	64.2	74.1	94.0	82.8	89.5		
	0.270	0.771	0.599	0.559	0.731	0.701	0.681	0.920	0.813	0.899		

$$\boldsymbol{p}_k = SoftMax(\mathbf{W}_3 \boldsymbol{h}_k) \tag{4}$$

$$SoftMax(\boldsymbol{p}_k) = \frac{exp(\boldsymbol{p}_k)}{\sum_i exp(\boldsymbol{p}_{k,i})} \tag{5}$$

$\mathbf{W}_3$ is the weight of FC layer and $\boldsymbol{p}_k$ is the class probability distribution of image $\tilde{\mathbf{x}}_k$. $SoftMax$ is the normalized exponential function whose expression is Eq. 5. $\boldsymbol{p}_{k,i}$ means the probability of image $\tilde{\mathbf{x}}_k$ belonging to class i.

3 Results

In quantitative evaluation, we use the Kappa coefficient (Kappa) and overall accuracy (OA) as the overall evaluation indexes and the producer accuracy (PA), user accuracy (UA), and F1 score (F1) as the evaluation indexes for each category. Table 2 shows the evaluation result of two initialization strategy with different numbers of finetuning samples. When 100% finetuning samples are used, the SSL method gains a better result than SL. The Kappa and OA increase by 2.4% and 2.1%, respectively. According to F1, the SSL method achieves the best results in 8 out of 10 categories. For both SSL and SL models, forests and water have an F1 value of above 0.9, due to the simple texture. Residential zones, transportations, open spaces, constructions, and agricultural lands are also visually distinguishable, so their F1 values are over 0.75. However, commercial, institutional, and industrial zones with strong social attributes are visually ambiguous, which are difficult to be accurately classified them using the visual characteristics provided by remote sensing images, so their F1 values are relatively low.

Figure 5 shows the UFZ map predicted by two models. One model is initialized by SSL on the research region (SSL model) with 100% finetuning samples and another is randomly initialized (SL model) with 100% finetuning samples. We show four results in Figure 5, which intuitively demonstrate the superiority of the SSL model in UFZ classification. The comparison chart shows that SL is prone to misclassifying UFZs with visual homogeneity, such as open space, forest, residential zone and commercial zone. For example, in region A, the SSL model accurately identifies the area with forest trails as forest, while the SL model misclassifies it as an open space. A possible reason is that the SL model cannot distinguish between

FIGURE 7
Samples misclassified by the model pretrained on AID. The texts at the bottom of each subfigure are the ground truth and the prediction. For example, the image in the left upper corner is for transportation, but it was misclassified as agricultural land. Com: commercial, Res: residential, Ins: institutional, Ind: industrial, Tra: transportation, OS: open space, Con: construction, For: forest, Agr: agricultural.

forest trails and park trails when the samples are limited, while the SSL model can distinguish between the two using many unlabeled samples.

3.1 The advantages of SSL in UFZ classification

To compare the performance of SSL and SL in UFZ classification, we carry out a set of experiments using 10%, 20%, 40%, 80%, and 100% training samples for finetuning, separately. The overall results are shown in Figure 6. For detailed qualitative evaluation, please refer to Table 2. Compared with the randomly initialized model (SL-based model), the model pretrained *via* SSL gains better results. Following are the advantages of SSL:

1) Using the same number of training samples, the SSL-based model achieves higher Kappa and OA than the SL-based model, and the fewer the training samples the more obvious the advantage. When 100% training samples are used for finetuning, the values of Kappa and OA of the SSL-based model are 2.1% and 2.0% higher than those of the SL-based model, respectively. When the samples reduce to 10%, the correspondence is 11.8% and 10.0%.
2) The SSL-based model achieves results comparable to or better than that got by the SL-based model but uses fewer samples. When the SSL-based model uses 10% (Kappa: 0.642; OA: 69.3%)and 20% (Kappa: 0.727; OA: 76.4%)samples for finetuning, the results are better than that got by SL-based models using 20% (Kappa: 0.636; OA: 68.2%)and 40% (Kappa: 0.709; OA: 74.7%)samples, respectively.

3.2 Spatial patterns of the urban functional zones

As shown in the map in Figure 5, there are many institutional zones in downtown, like government buildings, universities, and research institutes, because this city is the cultural and political center of China. The residential zones rank the top ratio in the center city. In the suburb, there are large areas of forest, open space, and agricultural land. The construction regions between urban and suburban areas reflect the expansion of Beijing.

In this study, we analyze the spatial patterns of the UFZs in the research region. The location quotient (LQ) is used to evaluate the ratio of specialization of a region (Kolars and Haggett, 1967). LQ is calculated by Eq. 6, in which the area ratio of UFZ c in region r is the divided by area ratio of total UFZ c in the research region s.

$$LQ_c^r = \frac{s_c^r / s^r}{s_c / s} \tag{6}$$

When $LQ_c^r > 1.5$, UFZ c has a high superiority in region r; when LQ_c^r is between 1 and 1.5, UFZ c exceeds the average level in region r; if $LQ_c^r < 1$, UFZ c is below the average level in region r.

By calculating LQ, the development status of UFZs and the degree of function composite can be analyzed.

The LQ based on ring roads (from the 2nd ring road to the 6th ring road) and administrative divisions (Xicheng, Dongcheng, Haidian, Chaoyang, Shijing, and Fengtai) is calculated and shown in Table 3 and Table 4.

Commercial, residential, and institutional zones show different superiority in the regions divided by ring roads. The commercial zone shows high superiority inside the 5th ring road, the institutional zones are concentrated inside the 4th ring road, and the residential zones are prominent inside the 2nd and between the 3rd and 4th ring roads. Apart from the downtown, the superiority of the above functional zones reduces, and other UFZs increase.

From the perspective of administrative divisions, the commercial, residential, and institutional zones show superiority in the inner city (Xicheng and Dongcheng district). The forest shows superiority in the Haidian and Shijing districts, as they share the Western Hills National Forest Park. In Chaoyang and Fengtai districts, most kinds of UFZs are at the average level.

4 Discussion

4.1 The gap between benchmarks and practical application

As we mentioned in the Introduction, SSL has been investigated deeply using different data, but it is rarely used in practical applications like UFZ classification, and the gap between benchmarks and practical applications is also ignored. Here, we conducted an experiment, in which two ResNet50 models are pretrained by the sample generated from the research region and the sample in the AID (Xia et al., 2017) dataset *via* SSL separately, and finetuned by 100% annotated samples collected in the research region. Table 5 compares the performance of the two models.

Compared with the model pretrained on the AID, the model pretrained on the research region gains 15.9% and 12.0% higher Kappa and OA. The average F1 got by the research region based model was 19.2% higher than the other, with the maximum increase of 62% in commercial areas.

As shown in Figure 2, there are many patches in the dataset with incomplete objects, while in public datasets such as AID, samples are carefully selected that have higher visual discrimination and are easier for the model to capture its features. Figure 7 shows some samples misclassified by the AID pre-trained model. The objects are not in the center of the patches and some objects are incomplete. But they have been accurately classified by the research region pre-trained model. For example, airplanes are important objects for identifying the airport, but they are not at the center of patches in the practical samples, which leads misclassification. This conflicts the prior knowledge learned from the benchmark that objects used to determine the category of images should be in the image center.

5 Conclusion

Current SL-based UFZ classification methods require a lot of training samples, which are not easy to acquire. Thus, this study conducts research on UFZ classification based on SSL. We collect 7304 typical UFZ samples as the finetuning and testing data and map the UFZ distribution inside the 6th ring road in Beijing. The experiment result proves that SSL gains better classification results than SL when the same number of training data is used and achieves comparable results to SL using half of the training samples. However, the classification accuracy of commercial, institutional, and industrial zones is still unsatisfying due to visual ambiguity. In addition, the comparison experiment between the model pretrained on the research region and that pretrained on the benchmark demonstrates the difficulties in the practical application of SSL. The displacement and incompleteness of objects in real data impact the performance of SSL models.

In the future, we will use social sensing data like geo-tagged photos, taxi trajectories, and points of interest as supplementary information for UFZ classification.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

WL: Data curation, Software, Visualization, Writing-Original draft preparation. JQ: Investigation and Reviewing. HF: Conceptualization, Methodology, Supervision, Editing.

Funding

This work was supported by the Inner Mongolia Science & Technology Plan (2022YFSJ0014).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Bao, H., Ming, D., Guo, Y., Zhang, K., Zhou, K., and Du, S. (2020). DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. *Remote Sens. (Basel).* 12, 1088. doi:10.3390/rs12071088

Cao, R., Tu, W., Yang, C., Li, Q., Liu, J., Zhu, J., et al. (2020). Deep learning-based remote and social sensing data fusion for urban region function recognition. *ISPRS J. Photogramm. Remote Sens.* 163, 82–97. doi:10.1016/j.isprsjprs.2020.02.014

Castelluccio, M., Poggi, G., Sansone, C., and Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092* https://arxiv.org/abs/1508.00092 (Accessed August 01, 2015).

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in Proceedings of the International Conference on Machine Learning (ICML) Proceedings of Machine Learning Research. (PMLR), 1597–1607, July 2020. Available at: http://proceedings.mlr.press/v119/chen20j.html (Accessed June 5, 2021).

Chen, W., Huang, H., Dong, J., Zhang, Y., Tian, Y., and Yang, Z. (2018). Social functional mapping of urban green space using remote sensing and social sensing data. *ISPRS J. Photogramm. Remote Sens.* 146, 436–452. doi:10.1016/j.isprsjprs.2018.10.010

Cheng, G., Han, J., and Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* 105, 1865–1883. doi:10.1109/jproc.2017.2675998

Cheng, G., Xie, X., Han, J., Guo, L., and Xia, G.-S. (2020). Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 13, 3735–3756. doi:10.1109/JSTARS.2020.3005403

Cheng, G., Yang, C., Yao, X., Guo, L., and Han, J. (2018). When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* 56, 2811–2821. doi:10.1109/tgrs.2017.2783902

Dai, D., and Yang, W. (2010). Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geosci. Remote Sens. Lett.* 8, 173–176. doi:10.1109/lgrs.2010.2055033

Doersch, C., and Zisserman, A. (2017). "Multi-task self-supervised visual learning," in Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 2017 (IEEE). doi:10.1109/iccv.2017.226

Du, S., Du, S., Liu, B., and Zhang, X. (2019). Context-enabled extraction of large-scale urban functional zones from very-high-resolution images: A multiscale segmentation approach. *Remote Sens. (Basel).* 11, 1902. doi:10.3390/rs11161902

Du, S., Du, S., Liu, B., and Zhang, X. (2021). Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. *Remote Sens. Environ.* 261, 112480. doi:10.1016/j.rse.2021.112480

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 303–338. doi:10.1007/s11263-009-0275-4

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016, 770–778.

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 12, 2217–2226. doi:10.1109/JSTARS.2019.2918242

Hong, D., Hu, J., Yao, J., Chanussot, J., and Zhu, X. X. (2021). Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogrammetry Remote Sens.* 13, 68–80. doi:10.1016/j.isprsjprs.2021.05.011

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of the International Conference on Machine Learning, Guangzhou, China, July 2015.

Kang, Y., Cho, N., Yoon, J., Park, S., and Kim, J. (2021). Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos. *ISPRS Int. J. Geoinf.* 10, 137. doi:10.3390/ijgi10030137

Kolars, J., and Haggett, P. (1967). Locational Analysis in human geography. *Econ. Geogr.* 43, 276. doi:10.2307/143300

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in Proceedings of the Advances in neural information processing systems, Lake Tahoe, NV, USA., December 2012, 1097–1105.

Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., et al. (2022). Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 60, 1–14. doi:10.1109/TGRS.2022.3147513

Liu, B. H., Deng, Y. B., Li, M., Yang, J., and Liu, T. (2021). Classification schemes and identification methods for urban functional zone: A review of recent papers. *Appl. Sci. (Basel).* 11, 9968. doi:10.3390/app11219968

Liu, H. M., Xu, Y. Y., Tang, J. B., Deng, M., Huang, J. C., Yang, W. T., et al. (2020). Recognizing urban functional zones by a hierarchical fusion method considering landscape features and human activities. *Trans. Gis* 24, 1359–1381. doi:10.1111/tgis.12642

Liu, Q., Hang, R., Song, H., and Li, Z. (2017). Learning multiscale deep features for high-resolution satellite image scene classification. *IEEE Trans. Geosci. Remote Sens.* 56, 117–126. doi:10.1109/tgrs.2017.2743243

Lu, W., Tao, C., Li, H., Qi, J., and Li, Y. (2022). A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sens. Environ.* 270, 112830. doi:10.1016/j.rse.2021.112830

Luo, B., Jiang, S., and Zhang, L. (2013). Indexing of remote sensing images with different resolutions by multiple features. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 6, 1899–1912. doi:10.1109/JSTARS.2012.2228254

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., and Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS J. Photogramm. Remote Sens.* 130, 277–293. doi:10.1016/j.isprsjprs.2017.06.001

Schmarje, L., Santarossa, M., Schröder, S. M., and Koch, R. (2021). A survey on semi-self- and unsupervised learning for image classification. *IEEE Access* 9, 82146–82168. doi:10.1109/ACCESS.2021.3084358

Stojnic, V., and Risojevic, V. (2021). "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, June 2021 (IEEE), 1182–1191. doi:10.1109/cvprw53098.2021.00129

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, June 2016, 2818–2826.

Tao, C., Qi, J., Lu, W., Wang, H., and Li, H. (2020). Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geosci. Remote Sens. Lett.* 1, 1–5. doi:10.1109/LGRS.2020.3038420

Tao, C., Qi, J., Lu, W., Wang, H., and Li, H. (2022). Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geosci. Remote Sens. Lett.* 19, 1–5. doi:10.1109/lgrs.2020.3038420

Tao, C., Yin, Z., Zhu, Q., and Li, H. (2021). Remote sensing image intelligent interpretation: From supervised learning to self-supervised learning. *Acta Geod. Cartogr. Sinica* 50, 1122–1134.

Wang, Q., Liu, S., Chanussot, J., and Li, X. (2018). Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 57, 1155–1167. doi:10.1109/tgrs.2018.2864987

Wang, X., Zhang, S., Yu, Z., Feng, L., and Zhang, W. (2020). "Scale-equalizing pyramid convolution for object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, June 2020, 13359–13368.

Xia, G. S., Hu, J. W., Hu, F., Shi, B. G., Bai, X., Zhong, Y. F., et al. (2017). Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* 55, 3965–3981. doi:10.1109/tgrs.2017.2685945

Xu, S., Qing, L., Han, L., Liu, M., Peng, Y., and Shen, L. (2020). A new remote sensing images and point-of-interest fused (RPF) model for sensing urban functional regions. *Remote Sens. (Basel).* 12, 1032. doi:10.3390/rs12061032

Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S., and Xie, P. (2020). Transfer learning or self-supervised learning? A tale of two pretraining paradigms. *arXiv: 2007.04234 [cs, stat]*. Available at: http://arxiv.org/abs/2007.04234 (Accessed September 19, 2021).

Yin, J., Dong, J., Hamm, N. A. S., Li, Z., Wang, J., Xing, H., et al. (2021a). Integrating remote sensing and geospatial big data for urban land use mapping: A review. *Int. J. Appl. Earth Observation Geoinformation* 103, 102514. doi:10.1016/j.jag.2021.102514

Yin, J., Fu, P., Hamm, N. A. S., Li, Z., You, N., He, Y., et al. (2021b). Decision-level and feature-level integration of remote sensing and geospatial big data for urban land use mapping. *Remote Sens.* 13, 1579. doi:10.3390/rs13081579

Yu, Y., Li, X., and Liu, F. (2020). Attention GANs: Unsupervised deep feature learning for aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* 58, 519–531. doi:10.1109/tgrs.2019.2937830

Zhang, X., Du, S., and Wang, Q. (2017). Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS J. Photogramm. Remote Sens.* 132, 170–184. doi:10.1016/j.isprsjprs.2017.09.007

Zhang, X., Du, S., and Wang, Q. (2018). Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sens. Environ.* 212, 231–248. doi:10.1016/j.rse.2018.05.006

Zhang, X., Du, S., and Zheng, Z. (2020). Heuristic sample learning for complex urban scenes: Application to urban functional-zone mapping with VHR images and POI data. *ISPRS J. Photogramm. Remote Sens.* 161, 1–12. doi:10.1016/j.isprsjprs.2020.01.005

Zhao, Z., Luo, Z., Li, J., Chen, C., and Piao, Y. (2020). When self-supervised learning meets scene classification: Remote sensing scene classification based on a multitask learning framework. *Remote Sens.* 12, 3276. doi:10.3390/rs12203276

Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., and Hong, Z. (2020). SO–CNN based urban functional zone fine division with VHR remote sensing image. *Remote Sens. Environ.* 236, 111458. doi:10.1016/j.rse.2019.111458

Zhu, Q., Zhong, Y., and Zhang, L. (2014). "Multi-feature probability topic scene classifier for high spatial resolution remote sensing imagery," in Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Quebec City, QC, Canada, July 2014, 1–4. doi:10.1109/IGARSS.2014.6947071