Index Your Position: A Novel Self-Supervised Learning Method for Remote Sensing Images Semantic Segmentation

Dilxat Muhtar, Xueliang Zhang, *Member, IEEE*, and Pengfeng Xiao, *Senior Member, IEEE*

***Abstract*— Learning effective visual representations without human supervision is a critical problem for the task of semantic segmentation of remote sensing images (RSIs), where pixel-level annotations are difficult to obtain. Self-supervised learning (SSL), which learns useful representations by creating artificial supervised learning problems, has recently emerged as an effective method to learn from unlabelled data. Current SSL methods are generally trained on ImageNet through image-level prediction tasks. We argue that this is suboptimal for application in semantic segmentation of RSIs since it does not take into account spatial position information between objects, which is critical for the segmentation of RSIs characterized by multiobject. In this study, we propose a novel self-supervised dense representation learning method, IndexNet, for the semantic segmentation of RSIs. On the one hand, considering the multiobject characteristics of RSIs, IndexNet learns pixel-level representations by tracking object positions, while maintaining sensitivity to object position changes to ensure that no mismatches are caused. On the other hand, by combining image-level contrast and pixel-level contrast, IndexNet can learn spatiotemporal invariant features. Experimental results show that our method works better than ImageNet pretraining and outperforms state-of-the-art (SOTA) SSL methods. Code and pretrained models will be available at https://github.com/pUmpKin-Co/offical-IndexNet.**

***Index Terms*— Remote sensing images (RSIs), self-supervised learning (SSL), semantic segmentation.**

I. Introduction

SEMANTIC segmentation of remote sensing images (RSIs), which aims to assign a geographic label to every pixel in an image, is a pivotal task in a wide range of real-world applications, such as land cover mapping [1], [2], infrastructure management [3], [4], and precision agriculture [5], [6]. The complicated spectral responses of multiple objects caused by different sensors, angles, and weather make the semantic segmentation of RSIs particularly challenging. Thanks to the boom in deep learning research in recent years, the performance of semantic segmentation of RSIs has achieved great progress [7]–[9].

However, semantic segmentation requires the collection of pixel-level class labels, which is tedious and requires rich experience and sound geographic knowledge. Although there are many publicly available annotated datasets [10]–[12], RSIs vary significantly in time and location, existing labeled data is merely an interception of the images and gathering a large number of annotated samples with an exceptionally high richness that encompass global regions, multiresolution, multiseason, and multispectral is difficult. One way to solve this problem is using transfer learning [13] to transfer the knowledge learned from a larger domain to improve performance on the target domain or reduce reliance on labeled samples. The most widely used transfer learning method for RSIs semantic segmentation is based on ImageNet [14]. However, this transfer learning strategy does not enhance performance significantly and does not make use of the enormous amount of unlabelled data.

The introduction of self-supervised learning (SSL) addresses this issue and makes use of vast amounts of unlabelled data. SSL methods can first learn useful representations from unlabelled source data by solving predesigned tasks (called pretext tasks) and then transfer them to target tasks (such as semantic segmentation). This possibility of using unlabelled images for representation learning has attracted considerable attention, resulting in substantial progress in SSL [15]–[19]. The self-supervision that guides representation learning in current methods is based on the image-level comparison. The latent prior of this learning pipeline is that different views (crops) of the same image correspond to the same object, as shown in Fig. 1(a). However, RSIs typically represent a wide spatial range due to overhead imaging, which leads to the possibility of containing different objects in one image. As a result, different random crops may correspond to different objects, as shown in Fig. 1(b) and (c). Moreover, current SSL methods are mainly designed for the image classification task that only requires image-level representation. However, semantic segmentation requires both image-level and pixel-level representations to yield promising results. Although a growing number of studies [20]–[23] investigate SSL pretraining for dense prediction tasks, these methods do not account for the unique

Manuscript received February 8, 2022; revised April 9, 2022; accepted May 16, 2022. Date of publication May 25, 2022; date of current version June 9, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297 and Grant 41871235, in part by the Fundamental Research Funds for the Central Universities under Grant 020914380095, and in part by the High-Level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China. *(Corresponding author: Xueliang Zhang.)*

The authors are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, and the Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: 181830041@smail.nju.edu.cn; zxl@nju.edu.cn; xiaopf@nju.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3177770

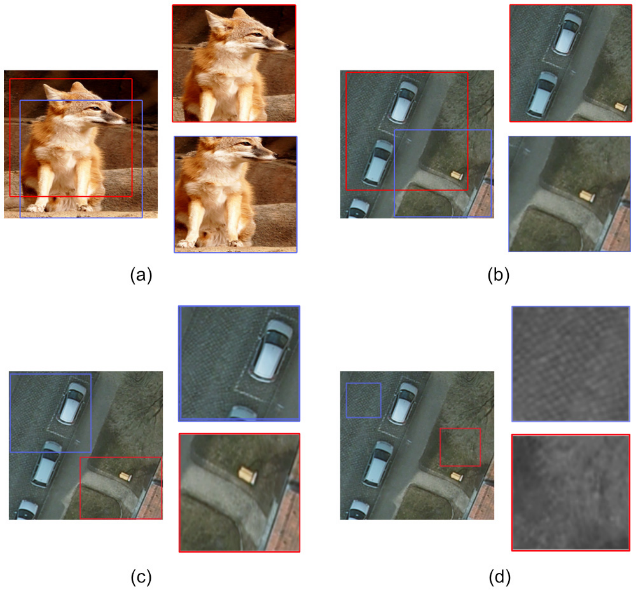

Fig. 1. **Random cropped views on (a) ImageNet and (b)–(d) Potsdam datasets**. (a) On ImageNet, different cropped views correspond to the same object. (b) However, on Potsdam, each image contains multiobject, thus different crops correspond to different objects. (c) Furthermore, if spatial position is not taken into account, two crops may not contain the same object at all, and (d) the selection of corresponding pixels based entirely on similarity will result in mismatches due to spectral transformation.

characteristics of RSIs. For example, DenseCL [20] and Self-EMD [23] learn pixel-level representation by matching pixels based on feature similarity, but these approaches are prone to mismatching corresponding pixels due to the complex and varied spectral response of objects in RSIs and the artificial augmentation of images during pretraining [see Fig. 1(d)]. DetCon [21] requires additional segmentation of images for SSL pretraining, which is time-consuming and not suitable for practical application of RSIs.

In this study, we propose a novel SSL method, IndexNet, for the semantic segmentation of RSIs. First, we introduce Index Contrast to account for the wide range of spectral responses of objects in RSIs, which allows our model to learn pixel-level representations and preserve the spatial position information to make the learned representations variant to different objects. Second, considering the substantial temporal variance in RSIs and the need for global information for semantic segmentation tasks, we combine the Instance Contrast method BYOL [18] with Index Contrast to learn both pixel-level and image-level spatiotemporal invariant representations. In general, we make the following contributions.

1) We show that, despite the complicated interactions among multiobject in RSIs, SSL can learn useful representations for semantic segmentation of RSIs using unlabelled data with careful design. Since we have effortless access to a large number of unlabelled RSIs, this learning method has great potential for promoting RSI semantic segmentation and its related applications.
2) Considering the multiobject characteristics of RSIs and the complicated interactions among objects, we introduce the Index Contrast mechanism, which makes it possible to keep invariant to spatial transformation in the process of learning pixel-level representation. When combined Index Contrast with Instance Contrast, our network can learn both pixel-level and image-level spatiotemporal invariant representations.
3) We evaluated IndexNet on two public datasets. When evaluated on the Potsdam dataset, the proposed IndexNet improves mIoU by 3% compared to ImageNet pretrained model and 1% compared to other SSL methods. Using only 10% of the labels, our pretrained model can achieve the equivalent accuracy as the randomly initialized model trained with the full labels. On the LoveDA dataset, IndexNet improves mIoU by 9% compared to the ImageNet pretraining method and outperforms other SSL methods.

II. Related Works

A. Semantic Segmentation

Since a fully convolutional network (FCN) [24] was proposed and achieved great success, deep learning methods have dominated the field of semantic segmentation. Several subsequent works [25], [26] were built upon the same idea of using the encoder–decoder architecture. Chen *et al.* [27] introduced the use of atrous convolutions to retain the receptive field of view and enhance the performance of the network. Later, various works focused on the aggregating long-range context in the final feature map [13], [28], [29] and using attention mechanisms to consider the relationship between the pixels [30]–[32].

At the same time, numerous segmentation networks for RSIs have been proposed [8], [9], [33], [34]. By considering variable scales and the hierarchical structure of RSIs, these methods have achieved better performance on this specific task.

B. Self-Supervised Learning

Initially, most SSL methods were based on pretext tasks, including colorization [35], inpainting [36], denoising [37], solving jigsaw puzzles [38], and predicting orientation [39].

Recently, the contrastive learning method has drawn much attention and achieved state-of-the-art (SOTA) performances. The main idea of contrastive learning is to maximize the similarity of different augmented views of the same image and pull the views from different images apart. SimCLR [15] was the first to show that SOTA performance can be achieved by contrastive learning without a memory bank. MoCo [16] proposed using a momentum-encoder and a queue of negative samples to release the requirement of large batch size. BYOL [18] and SimSiam [40] aimed to learn representation without negative samples. Barlow-Twins [17] learned to make the cross-relation matrix between two augmented views of the same image as close as possible to an identity matrix. SwAV [19] incorporated clustering into the contrastive learning framework by computing the assignment from one view and predicting it from another view.

While these pretraining methods achieve promising results when transferred to the classification task, the progress on transfer performance for dense prediction tasks is limited [20].

Several attempts have been made to solve such a problem. DenseCL [20], PixPro [22], and Self-EMD [23] shift image-level representation to pixel-level representation by maximizing the similarity of corresponding pixels of different views. DetCon [21] uses the classic segmentation method to generate masks before selecting positive and negative samples at the pixel level for contrastive learning. Our work is similar to Self-EMD and DenseCL, except that we employ the Index Contrast mechanism to pick the corresponding pixels, which reduces the mismatches caused by matching pixels according to feature similarity. Moreover, compared to DetCon, our pretraining process is performed end-to-end, making it more computationally efficient and well-suited for the semantic segmentation task of RSIs.

C. SSL in Remote Sensing

Recently, researchers have also attempted to apply SSL to the analysis of RSIs. For example, Vincenzi *et al.* [41] proposed a method similar to image colorization to learn meaningful representations. Stojnic and Risojevic [42] evaluated the applicability of contrastive learning in RSIs classification and showed that self-supervised models trained on RSIs gave better results than supervised models trained on ImageNet. Tao *et al.* [43] analyzed the impacts of the choice of self-supervised signals, the domain difference between the source and target datasets, and the amount of pretraining data in SSL on the RSIs scene classification task. Like MoCo [16], SauMoCo [44] also used augmentation criteria with a contrastive loss formulation and a momentum updated-based optimization to exploit the semantic similarities and inherent diversity within land cover concepts. Considering that remote sensing data can provide repeated observations of the same location over time, [45] leveraged spatially aligned images over time to contrast temporal positive pairs and introduced a pretext task of predicting where in the world an image comes from. SeCo [46] also used images from the same location at different temporal phases as positive pairs, but the difference was that SeCo used the idea of multiple embedding subspace to learn common representations that encode the different variances and invariances. MATTER [47] proposed a texture refinement network to amplify low-level features and adapted residual cluster learning to learn material and texture-based representations for RSIs.

The above works were developed to learn global image representations that are most suitable for image-level tasks. Our work focuses on learning both image-level and pixel-level representations for the semantic segmentation task of RSIs.

III. Method

IndexNet consists of the Index Contrast branch and the Instance Contrast branch, which learns pixel-level and image-level representations by minimizing the loss functions $\mathcal{L}_I$ and $\mathcal{L}_P$, respectively, as illustrated in Fig. 2. In this section, we start by briefly introducing the learning framework of BYOL [18], the SOTA SSL method on which our study is based. Then, we describe the innovation we made when applying BYOL to the semantic segmentation of RSIs.

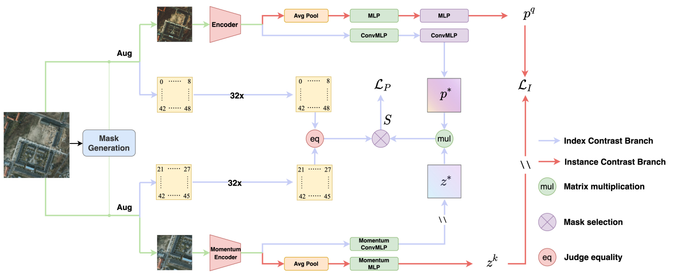

Fig. 2. **IndexNet architecture**. IndexNet has two branches, the Instance c Contrast branch and the Index Contrast branch, which learn image-level and pixel-level representations, respectively.

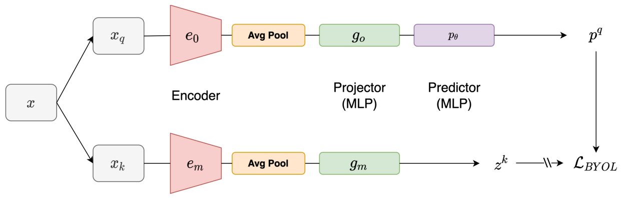

Fig. 3. **BYOL architecture**. BYOL learns image-level representations by Instance Contrast.

A. Framework of BYOL

As shown in Fig. 3, for each image x in a given unlabelled dataset, BYOL first generates two views x^q, x^k of x by random data augmentation. The two views x^q, x^k are passed through the online encoder e_o and the momentum encoder e_m, followed by global average pooling and projector g_o, g_m, respectively, to encode each view with embedding vectors z^q, z^k

$$z^q = g_o\big(\text{AvgPool}\big(e_o\big(x^q\big)\big)\big) \quad (1)$$
$$z^k = g_m\big(\text{AvgPool}\big(e_m\big(x^k\big)\big)\big). \quad (2)$$

In BYOL, e_o is a standard ResNet50 [48] that removes the final linear layer, and g_o is a one hidden layer multilayer perception (MLP). e_m and g_m have the same architecture as e_o and g_o except that the weights of the e_m and g_m are exponential moving averages of the online encoder e_o and the online projector g_o. That is, given a momentum rate $m \in [0, 1]$, e_m, g_m are updated by the following rule:

$$e_m = me_m + (1-m)e_o \quad (3)$$
$$g_m = mg_o + (1-m)g_o. \quad (4)$$

Then, z^q is fed to predictor p_θ with the same architecture as projector g_o and output p^q. Finally, BYOL ℓ_2−normalizes p^q and z^k to compute the final mean squared error (mse) loss

$$\mathcal{L}_{\text{BYOL}}(p^q, z^k) \triangleq ||p^q - z^k||_2^2 = 2 - 2\cdot < p^q, z^k > \quad (5)$$

B. Index Contrast

Although the existing SSL learning paradigm is able to learn powerful image-level representations, it is insufficient for semantic segmentation task that requires both image-level and pixel-level representations. Furthermore, current popular SSL approaches assume that the two different augmented images are from the same object, but when training on the unlabelled

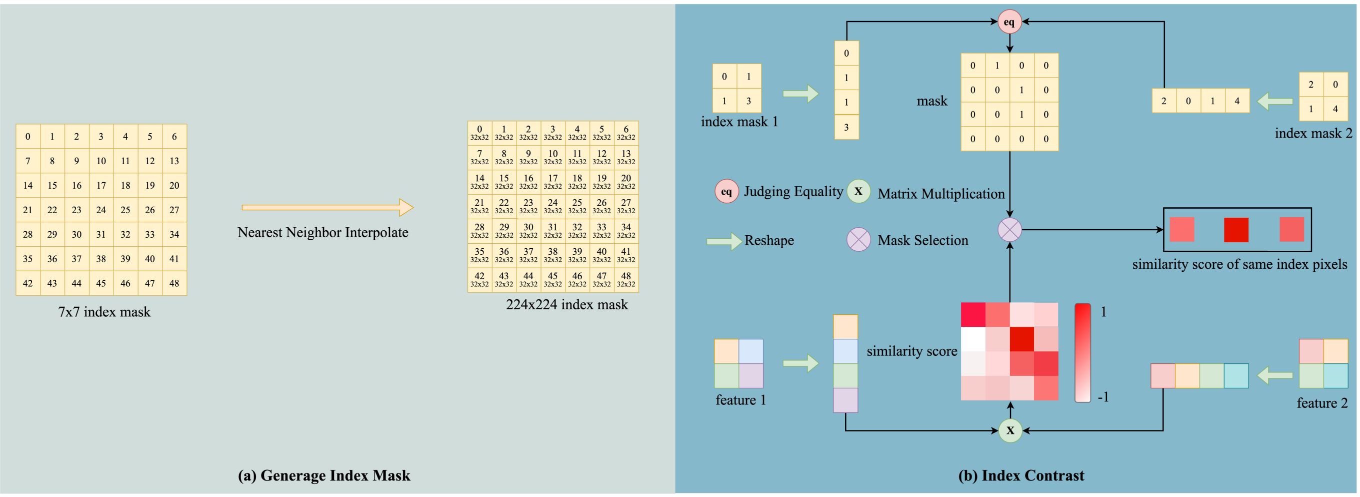

Fig. 4. **Illustration of Index Contrast branch. (a) Index mask generation procedure**, in which we construct a 7 × 7 matrix, assign a unique value to each pixel, and then interpolate the index mask to 224 × 224. **(b) Index Contrast.** We choose the similarity scores between pixels with the same index. For simplicity, we show the case where the feature map spatial dimension is 2 × 2.

TABLE I
PARAMETERS USED TO GENERATE IMAGE AUGMENTATIONS

Parameter	Value
Random crop probability	1.0
Horizontal Flip probability	0.5
Rotate probability	0.3
Color jittering probability	0.8
Brightness adjustment max intensity	0.4
Contrast adjustment max intensity	0.4
Saturation adjustment max intensity	0.4
Hue adjustment max intensity	0.2
Grayscaling probability	0.2
Gaussian blurring probability	0.5
Gaussian noise probability	0.6

RSIs, the noise from multiobject may inhibit the learning process, resulting in poor segmentation results.

To address this issue, it is natural to remove the global pooling layer and maximize the similarity of the corresponding pixels in the two views. However, the spatial location of the pixels has changed after the artificial augmentation, making it difficult to locate the corresponding pixels in the two views. Inspired by the semantic segmentation ground truth mask, we propose to index the pixel position to generate an index mask for each view. Considering that the final feature map has a spatial dimension of 7 × 7, we generate an index mask of size 7 × 7 and assign a value between 0 and 48 for each element. After creating the index mask, we use the nearest neighbor interpolation method to interpolate it up to the input image size (e.g., 224 × 224), resulting in each 32 × 32 image block corresponding to the same index value [see Fig. 4(a)]. Then, we transform each index mask using the same spatial transformations as those for the underlying views, resulting in two index masks M_1, M_2. Finally, we interpolate M_1 and M_2 to the same size as the feature map (7×7) using the nearest neighbor interpolation method.

For the feature maps $p^*, z^* \in \mathbb{R}^{H\times W\times C}$, where H and W denote the spatial size of the feature map (7 × 7) and C is the number of channels, we first ℓ_2-normalize them in the channel dimension and calculate the cosine similarity between the pixels of p^* and z^*. Then, we check one by one whether the index values between the pixels in M_1, M_2 are the same to generate a binary mask. Finally, according to this binary mask, we obtain $S \in \mathbb{R}^N$, which is the similarity score of the corresponding pixels, and N is the number of corresponding pixel pairs [see Fig. 4(b)].

C. Data Augmentation

SimCLR [15] demonstrates that data augmentation plays an important role in contrastive learning. In this study, we follow the data-augmentation strategy of GLCNet [49]. As shown in Table I, we use spatial transformations (random cropping, flipping, and rotating) and spectral transformations (color distortion, Gaussian blur, random noise, and random grayscaling). In addition, we index each pixel's position to build an index mask, which segments the image into separate regions (see Section III-B). In any case, we use the identical spatial transformations for each index mask as we do for its counterpart view to track the spatial position of different objects and find the corresponding pixels at the final step. This augmentation strategy encourages IndexNet to learn representations that are invariant to both spatial and spectral transformations.

D. Combine Instance Contrast and Index Contrast

1) Instance Contrast: The Instance Contrast branch is the same as the standard BYOL [18] learning pipeline. Specifically, for embedding vectors z^q, z^k in (1) and (2), we feed them into the regular MLP predictor, which outputs predictions p^q, p^k. Finally, we compute Instance Contrast loss

$$\mathcal{L}_I = \frac{1}{2}[\mathcal{L}_{\text{BYOL}}(p^q, z^k) + \mathcal{L}_{\text{BYOL}}(p^k, z^q)]. \tag{6}$$

2) Index Contrast: For the Index Contrast branch, we replace the projector and predictor architecture with pixel-wise MLP, which is implemented by a 1 × 1 convolution layer. In this way, we keep the spatial dimension of the feature map. The similarity scores S of the corresponding pixels are then obtained using the Index Contrast procedure described in Section III-B and are displayed in Fig. 4. Finally, we compute the index contrast loss as follows:

$$\mathcal{L}_P = \frac{1}{N}\sum_{i=1}^{N}(2 - 2 \cdot s_i) \tag{7}$$

where N is the number of selected similarity scores and $s_i \in S$. The final loss is defined as follows:

$$\mathcal{L} = \alpha\mathcal{L}_I + \beta\mathcal{L}_P \tag{8}$$

where α and β act as the weight to balance the two branches. α and β are set to 1 by default, the effectiveness of which is validated by experiments in Section IV-C1.

IV. Experiments

We first present implementation details in our training framework. Then we compare our method with the ImageNet pretraining baseline and other SSL methods for semantic segmentation of RSIs.

A. Implementations Details

1) Data Description: We use the International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam dataset[1] and the LoveDA dataset [12] for pretraining and fine-tuning. The details of the two datasets and data preprocessing are explained below.

a) ISPRS Potsdam dataset: It consists of 38 tiles of the same size of 6000 × 6000 pixels with a spatial resolution of 5 cm and four spectral bands: red, blue, green, and near-infrared. The dataset has been manually labeled into six categories: low vegetation, tree, building, impervious surface, car, and clutter. We crop these 38 tiles of images to 256 × 256 without overlap, yielding a total of 20 101 images. All of the cropped images are used for pretraining. We followed the train–test split strategy of [50] and used 23 tiles of images for fine-tuning and 14 tiles of images for testing the performance.

b) LoveDA dataset: It contains 5987 high spatial resolution (0.3 m) RSIs with a size of 1024 × 1024 pixels and three spectral bands (red, blue, and green) from Nanjing, Changzhou, and Wuhan in China. This dataset is separated into two parts: urban and rural. It contains seven categories: background, building, road, water, barren, forest, and agriculture. Our experiments take this urban part and crop these images to 256 × 256 pixels without overlap, resulting in 53 328 images. All of these cropped images are used for pretraining, while the training set (21 856 images) is used for fine-tuning, and the validation set (15 872 images) is used for testing since the labels for the test set are not available.

2) Pretraining Setting: For an image size of 256 × 256, we resize it to 224×224 using the bicubic interpolation method to align with the index mask. Using the setting as in [18], we choose ResNet50 [48] as our backbone in the two encoders. The projector of a standard MLP has a hidden dimension of 4096 and an output dimension of 256, whereas the predictor has a hidden dimension of 512 and an output dimension of 256. As mentioned in Section III-D, we replace the linear layer in the standard MLP with 1 × 1 convolution layer for the pixel-wise MLP (ConvMLP in Fig. 2). The exponential moving average parameter m in (3) and (4) starts from 0.996 and increases to 1 by the cosine annealing scheduler during training.

We use the stochastic gradient descent (SGD) optimizer with a weight decay of 1.5×10^{-6} for both the Potsdam and LoveDA datasets. We train the model with mixed precision on a single NVIDIA 2080Ti GPU with a batch size of 64. The learning rate is set to 0.03 and decreased to 5×10^{-7} following the cosine annealing scheduler.

To evaluate the performance of IndexNet, we use the common ImageNet pretrained model as the baseline, and we compare IndexNet with four SOTA instance-wise SSL methods: BYOL [18], Barlow-Tinws [17], MoCo-v2 [51], and SwAV [19]. To demonstrate that our proposed method is suitable for semantic segmentation of RSIs, we also compare IndexNet to SSL methods designed for dense prediction tasks: DenseCL [20] and PixPro [22]. We use the solo-learn library [52] to make these methods work. The hyperparameters in each method are the same as in the original paper, except that we linearly adjust the learning rate based on the reimplement batch size because of GPU limitations.

3) Fine-Tuning Setting: DeepLabV3+ [28] is used for downstream semantic segmentation tasks on both the Potsdam and LoveDA datasets with a batch size of 32. To evaluate the performance of the pretraining model, we not only use the regular fine-tuning strategy, which updates both the backbone and the decoder at the same time, but we also fine-tune with the frozen backbone, which we believe better evaluates the representation learned by the model during pretraining and can exclude the effects of hyperparameters in fine-tuning.

a) Fine-tuning: We use the SGD optimizer with a weight decay of 1×10^{-4} for both the Potsdam and the LoveDA datasets. The backbone learning rate for the Potsdam dataset is 0.001, while the decoder learning rate is 0.01. For the LoveDA dataset, the learning rates of both the backbone and the decoder are 0.01.

b) Freeze backbone fine-tuning: In this case, we freeze the backbone weights and train the decoder using SGD optimizer with a learning rate of 0.01 for the Potsdam dataset and 0.03 for the LoveDA dataset.

B. Experimental Results

1) Potsdam Dataset: We begin by evaluating the performance of the proposed IndexNet pretraining backbone on Potsdam with varying label sizes. As shown in Table II, pretraining with unlabeled data, IndexNet is able to learn useful representations and perform better than the ImageNet pretrained model when transferred to the semantic segmentation task with the same label size. Moreover, fine-tuning using the IndexNet pretrained model outperforms training with all labels with random initialization when only 10% of the labels are used.

In addition, we also compare IndexNet with other SSL methods. Tables III and IV show that our method outperforms

[1] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

TABLE II

COMPARED WITH RANDOM INITIALIZATION AND THE IMAGENET PRETRAINED MODEL. AFTER PRETRAINING INDEXNET ON POTSDAM FOR 400 EPOCHS, WE FINE-TUNE FOR 100 EPOCHS AND EVALUATE THE RESULTS ON THE TEST SET

Method	Label Size	OA(%)	mIoU(%)	Kappa
ImageNet	1%	71.19	49.59	0.6204
IndexNet	1%	**74.89**	**52.80**	**0.6699**
IndexNet	10%	**83.56**	**66.01**	**0.7874**
random init	100%	82.32	64.15	0.7682
ImageNet	100%	85.09	68.50	0.8043
IndexNet	100%	**86.87**	**71.07**	**0.8276**

TABLE III

FINE-TUNING ON THE POTSDAM DATASET USING 1% AND 10% LABELS. RESULTS ARE OBTAINED BY PRETRAINING 100 EPOCHS ON THE POTSDAM DATASET AND FINE-TUNING 50 EPOCHS

Method	1% Label			10% Label		
	OA(%)	mIoU(%)	Kappa	OA(%)	mIoU(%)	Kappa
SwAV	69.85	47.50	0.6045	80.84	62.40	0.7482
Barlow-Twins	69.88	46.96	0.6058	80.74	62.10	0.7467
MoCo-v2	72.28	50.39	0.6361	82.13	63.58	0.7657
BYOL	71.73	49.78	0.6291	81.02	62.18	0.7500
DenseCL	73.06	49.96	0.6452	82.51	63.70	0.7707
PixPro	66.95	43.31	0.5655	81.18	62.73	0.7532
IndexNet	**73.42**	**52.22**	**0.6526**	**82.96**	**64.60**	**0.7763**

TABLE IV

SEMANTIC SEGMENTATION ON THE POTSDAM DATASET. THE RESULTS ARE OBTAINED BY PRETRAINING 100 EPOCHS FOR ALL SSL METHODS. THE EPOCHS FOR FINE-TUNING AND FREEZE BACKBONE FINE-TUNING ARE 50 AND 20, RESPECTIVELY

	Fine-tuning			Freeze Backbone Fine-tuning		
	OA(%)	mIoU(%)	Kappa	OA(%)	mIoU(%)	Kappa
random init	70.93	46.63	0.6047	-	-	-
ImageNet	84.61	67.66	0.7979	82.41	64.64	0.7696
SwAV	85.02	68.44	0.8035	82.32	63.83	0.7679
Barlow-Twins	85.04	67.86	0.8035	82.08	63.15	0.7650
MoCo-v2	85.81	69.60	0.8144	82.96	64.80	0.7763
BYOL	85.25	68.27	0.8062	82.51	64.77	0.7714
DenseCL	85.35	67.45	0.8076	82.65	64.73	0.7733
PixPro	85.72	69.53	0.8126	77.60	57.09	0.7072
IndexNet	**86.45**	**70.71**	**0.8223**	**83.22**	**65.33**	**0.7806**

all other methods in downstream semantic segmentation tasks, whether fine-tuning with limited labels or full labels. Furthermore, the pretrained model obtained from the IndexNet pretraining method outperforms all other SSL methods at different epochs, as shown in Fig. 5, and achieves higher accuracy at the beginning of fine-tuning, implying that the model learns richer representation in the pretraining stage. We also calculate the single-class intersection over union (IoU) to see if our method has an advantage in each class, and the results are shown in Fig. 6, where it is clear that our method outperforms other SSL methods in all classes. Meanwhile, we show some visualization results in Fig. 7, where our method is relatively better.

To further show that the IndexNet pretrained model learns richer representation, in Table IV, we freeze the backbone (i.e., do not update the parameters) and only train the decoder. From the results, we can see that IndexNet also outperforms the other methods in this setting.

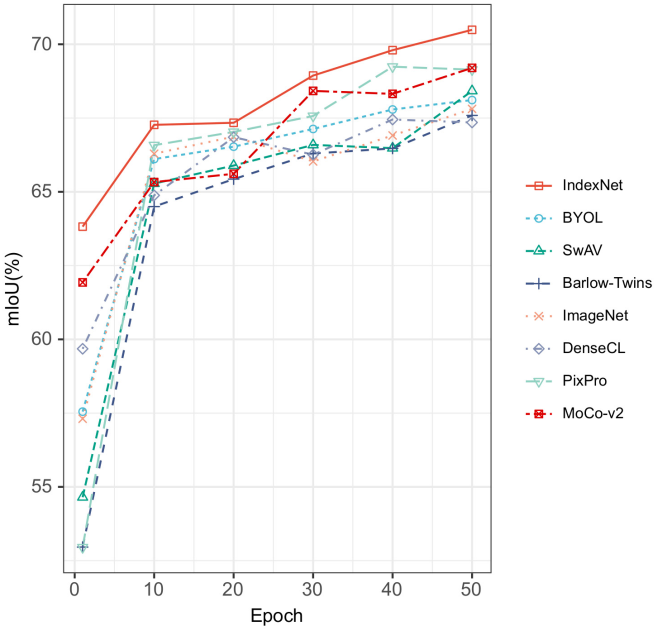

Fig. 5. **Fine-tuning on the Potsdam dataset with different epochs.** IndexNet outperforms other SSL methods at all epochs.

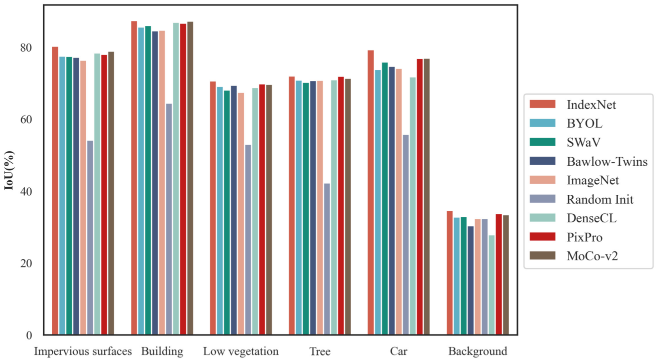

Fig. 6. **Class IoU of different methods on the Potsdam dataset.**

2) LoveDA Dataset: We also evaluate the performance of the IndexNet pretrained model on the LoveDA dataset. The semantic segmentation result is provided in Table V. We can observe that the segmentation results obtained using the IndexNet pretraining method are also better than those obtained using ImageNet and other SSL methods, regardless of whether the backbone is frozen. Moreover, we observe that the loss of IndexNet continues to decrease after 100 epochs of pretraining, thus, we pretrain for a longer time (400 epochs), and the segmentation results are further improved (see Table V).

3) Cross-Domain Evaluation: We compare the performance of models pretrained by different SSL methods when the datasets used for pretraining and fine-tuning are different to evaluate the impact of domain differences on the performance of the self-supervised pretraining model. The results are shown in Table VI. Our method outperforms all other SSL methods, which indicates that the IndexNet pretrained model is more robust to domain differences. More excitingly, even though the data used for pretraining in Potsdam (~20 k) and LoveDA

TABLE V

SEMANTIC SEGMENTATION ON THE LOVEDA DATASET. IF NOT SPECIFIED, ALL SSL METHODS ARE PRETRAINED WITH 100 EPOCHS. THE EPOCHS FOR FINE-TUNING AND FREEZE BACKBONE FINE-TUNING ARE 50 AND 20, RESPECTIVELY

	Fine-tuning			Freeze Backbone Fine-tuning		
	OA(%)	mIoU(%)	Kappa	OA(%)	mIoU(%)	Kappa
random init	58.67	25.32	0.4152	-	-	-
ImageNet	68.24	35.80	0.5483	62.86	35.81	0.4734
SwAV	68.93	41.45	0.5529	63.11	37.66	0.4807
Barlow-Twins	68.19	40.07	0.5401	64.59	38.22	0.4903
MoCo-v2	69.43	40.83	0.5541	66.52	39.35	0.5142
BYOL	66.78	39.97	0.5244	66.43	40.43	0.5147
DenseCL	68.60	43.91	0.5543	65.18	39.58	0.5045
PixPro	69.62	41.60	0.5533	64.59	38.03	0.4952
IndexNet	**71.79**	**44.18**	**0.5859**	**68.51**	**42.07**	**0.5245**
IndexNet(400ep)	**71.98**	**44.98**	**0.5879**	**71.67**	**43.20**	**0.5784**

TABLE VI

CROSS-DOMAIN EVALUATION. ALL RESULTS ARE OBTAINED BY PRETRAINING 100 EPOCHS AND FINE-TUNING 50 EPOCHS ON THE TARGET DATASET

Method	Potsdam→LoveDA			LoveDA→Potsdam		
	OA(%)	mIoU(%)	Kappa	OA(%)	mIoU(%)	Kappa
SwAV	65.20	39.39	0.5093	83.36	65.07	0.7813
Barlow-Twins	64.61	37.39	0.4961	83.52	65.47	0.7831
MoCo-v2	65.16	39.44	0.4934	84.88	66.98	0.8011
BYOL	64.83	38.58	0.5048	83.91	66.71	0.7885
DenseCL	67.73	38.70	0.5314	84.11	66.17	0.7914
PixPro	66.86	36.62	0.5142	83.64	65.16	0.7854
IndexNet	**67.96**	**40.82**	**0.5256**	**85.03**	**67.85**	**0.8026**

TABLE VII

COMPARISON OF SECO AND INDEXNET PRETRAINED MODELS ON THE POTSDAM SEMANTIC SEGMENTATION TASK

Method	OA(%)	mIoU(%)	Kappa
IndexNet	**84.22**	**66.99**	**0.7928**
SeCo	83.40	65.36	0.7818

(∼50 k) is smaller compared to ImageNet (∼1 M), the cross-domain performance of IndexNet is comparable to that of supervised pretraining on ImageNet (see Tables IV–VI).

4) SeCo Dataset: To further demonstrate the effectiveness of IndexNet for downstream semantic segmentation tasks, we compared IndexNet with SeCo [46]. We download the SeCo pretrained model[2] and train IndexNet on the SeCo-100K dataset [46] for 200 epochs following the SeCo implementation. We then fine-tune them at Potsdam for 50 epochs. In this setting, IndexNet also outperforms SeCo (see Table VII). However, we find that, although the SeCo-100K dataset is larger than the LoveDA dataset, the cross-domain transfer performance of the model pretrained on the SeCo-100K dataset was slightly worse than pretrained on the LoveDA dataset. We attribute this to the large domain gap between the Potsdam and SeCo-100K datasets caused by the large resolution differences.

C. Ablation

In this section, we conduct several experiments to show how each component contributes to IndexNet. We report ablation study by pretraining on Potsdam for 100 epochs and fine-tuning on Potsdam for 50 epochs.

TABLE VIII

IMPACT OF DIFFERENT BRANCH WEIGHTS

α	β	OA(%)	mIoU(%)	Kappa
1	0	85.25	68.27	0.8062
0	1	85.51	69.15	0.8205
1	1	86.45	70.71	0.8223
0.5	1	**86.52**	70.57	**0.8230**
1	0.5	86.49	**70.99**	0.8224

TABLE IX

IMPACT OF DIFFERENT LOSSES. WE ONLY USE THE INDEX CONTRAST BRANCH FOR COMPARISON WITH [21]

Loss	OA(%)	mIoU(%)	Kappa
infoNCE	83.89	66.25	0.7880
MSE	**85.51**	**69.15**	**0.8205**

1) Weights of the Two Branches: In this experiment, we examine how varying weights in (8) affect the results. First, to prove that combining the Instance Contrast and Index Contrast branches is effective, we choose BYOL [18] as our baseline, which only includes the Instance Contrast branch ($\alpha = 1, \beta = 0$). As shown in Table VIII, replacing Instance Contrast with Index Contrast ($\alpha = 0, \beta = 1$) improves segmentation accuracy slightly. When we combine the two branches, the segmentation accuracy of the model improves significantly, which further demonstrates that both image-level representation and pixel-level representation are necessary for the semantic segmentation task. Furthermore, we change the weights of the two branches to see if we could further improve. However, we found that this did not bring much improvement. Thus, we fix the weights of both branches to 1 by default.

2) Loss: We follow the implementation of [21] and consider pixels with different indices as negative samples to calculate infoNCE [53] loss. However, as shown in Table IX, the result is far worse than using mse loss. Based on this experiment, we can infer that in IndexNet, the index only serves to find pixels at the same position and cannot be used as a strategy for selecting negative samples.

3) Multilevel: Current SSL methods focus on learning discriminative features at the final layer. However, RSIs are characterized by hierarchical structure and most deep learning-based semantic segmentation methods try to extract features from the multilevel representation [7]–[9]. Thus, we attempt to apply the SSL method on different feature layers instead of only the final one on pyramids to investigate whether this approach boosts the performance. Specifically, for the Index Contrast branch, we downsample the index mask to the same size as the feature map (i.e., 56, 28, 14, and 7) at each level and perform index contrast as described in Section III-B. For the Instance Contrast branch, we pool the feature with average pooling and project it through an MLP with a hidden layer 2× feature dimension and an output dimension of 256. The results are shown in Table X, where there is a decrease

[2] https://github.com/ElementAI/seasonal-contrast

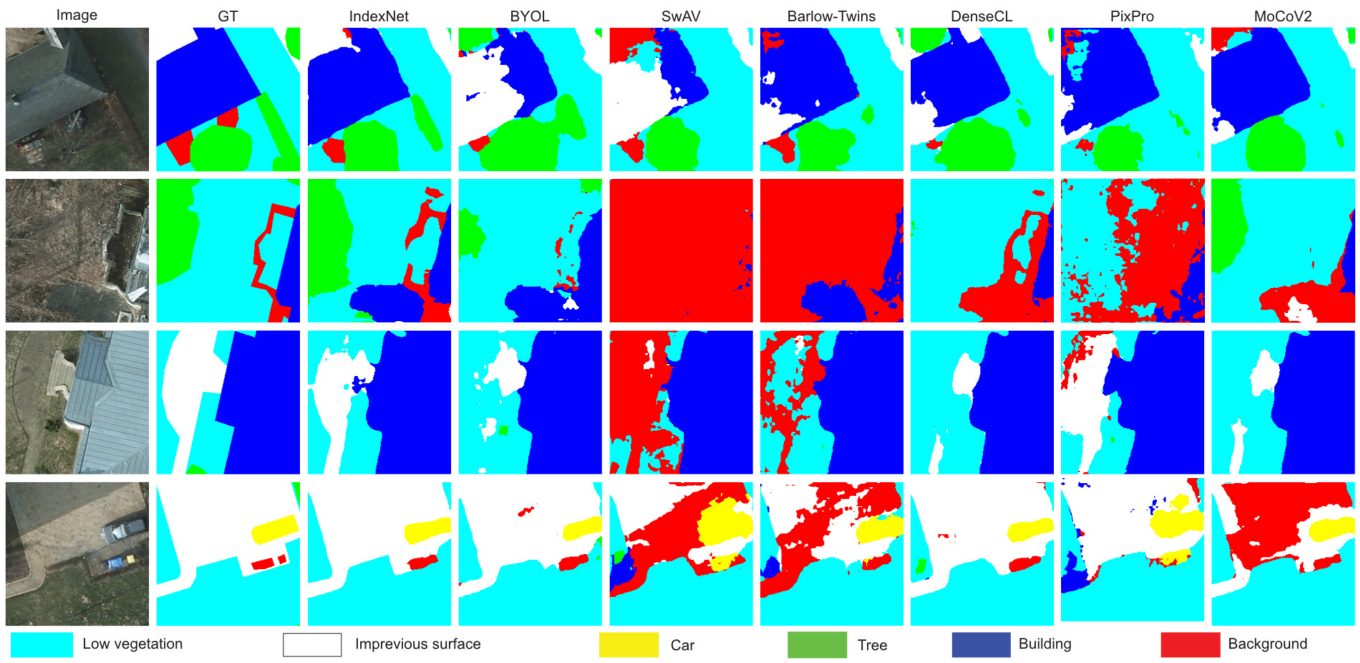

Fig. 7. **Visualization results of different methods on the Potsdam dataset**. GT refers to ground truth.

TABLE X
ABLATION RESULTS IN MULTILEVEL CONTRAST FOR DIFFERENT BRANCHES

Branch	Level	OA(%)	mIoU(%)	Kappa
Index Contrast	{4}	**85.51**	**69.15**	**0.8205**
Index Contrast	{1, 2, 3, 4}	84.92	68.10	0.8022
Two Branches	{4}	**86.45**	**70.71**	**0.8223**
Two Branches	{3, 4}	86.44	70.30	**0.8223**
Two Branches	{2, 3, 4}	86.05	69.81	0.8173
Two Branches	{1, 2, 3, 4}	85.97	69.71	0.8160

in accuracy regardless of which branch is used for multilevel contrast.

We interpret this phenomenon as follows: the shallow layers of CNN learn low-level features [54], which are important for every object, and if we contrast them to make them specific, the discriminative power of the model will decrease, while the deeper layer of CNN learns more specific features [54], and contrast them will increase the discriminative power of the model. This interpretation is further supported by Table X, which shows that SSL at higher levels gives better results after fine-tuning.

V. DISCUSSION

A. Why Index Contrast?

In this section, we provide a theoretical explanation for the design concept of Index Contrast to address the multiobject problem shown in Fig. 1 and confirm the explanation experimentally.

In the Index Contrast branch, we assign a unique index value to each 32 × 32 block in the image and filter similarity scores between two views with the same local index (see Fig. 4). As a consequence, even though the two views do not overlap [see Fig. 1(c)] or if the two views with different objects are highly similar after spectral transformation [see Fig. 1(d)], there will be no mismatch between pixels. Moreover, we assume that pixels with the same index correspond to the same object, while pixels with different indices correspond to different

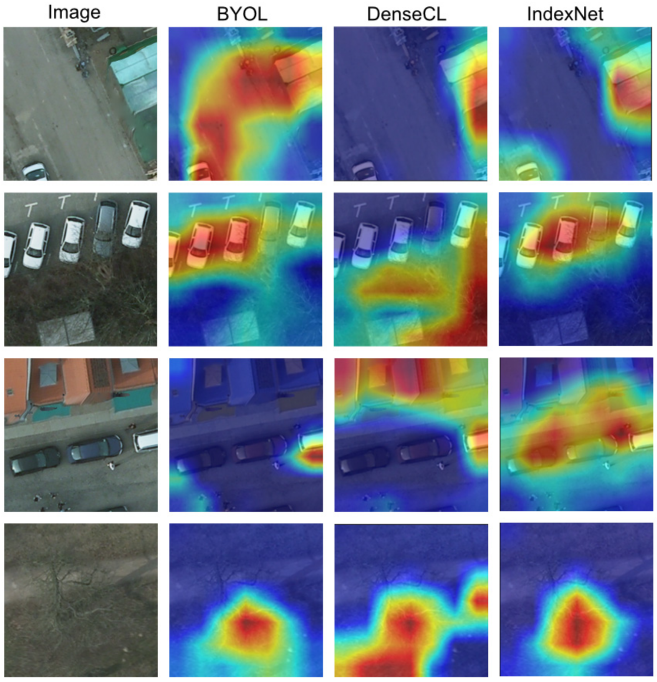

Fig. 8. **Comparison of the heatmap visualization under BYOL, DenseCL, and IndexNet by GradCam [55] on the Postsdam dataset.** Each heatmap is constructed by the last convolutional layer of the 100-epoch pretrained ResNet-50 model.

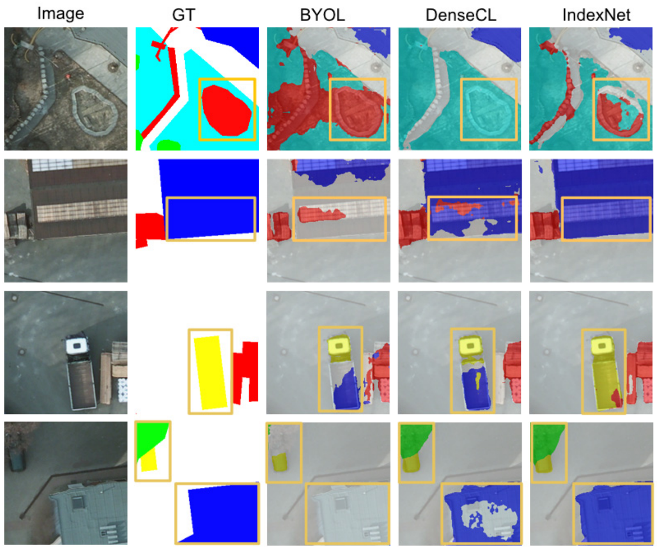

Fig. 9. **Comparison of semantic segmentation under BYOL, DenseCL, and IndexNet on the Potsdam dataset.** The results are produced by pretraining on the Potsdam dataset for 100 epochs and fine-tune for 50 epochs. We blend the raw image with the model prediction for better visualization. GT refers to ground truth.

objects. Hence, if there are overlapping areas in the two views [see Fig. 1(b)], we maximize the similarity between pixels belonging to the same object (i.e., same index). In this manner, we believe that the model is not only able to learn pixel-level representation, but also can make the learned representation variant for different objects, thus effectively solving the multiobject problem existing in RSIs. This is supported by Fig. 8, in which our IndexNet focuses more on region-level features than other methods, with each region containing only similar objects and the region boundaries being more accurate. We also visualize some results of different models on the Potsdam semantic segmentation task in Fig. 9. We can find that BYOL is more prone to misjudging the pixels' category as it only focuses on image-level representation. For example, in the second and fourth rows, BYOL incorrectly categorizes a large range of buildings. Despite the fact that DenseCL is devoted to extracting pixel-level representation, Fig. 9 shows that there are still some mistakes in pixel category assignment, particularly when two different categories in the image are highly similar (the first and second rows in Fig. 9). It has a lot to do with the fact that DenseCL relies on selecting pixels belonging to the same object based only on feature similarity, without considering the spatial location, which can easily lead to mismatches. With our IndexNet, the performance gets better in the segmentation task, proving the effectiveness of IndexNet for handling the multiobject problem in RSIs.

However, there are two scenarios in which the assumptions behind Index Contrast may not be satisfied. The first is that blocks with different index values may correspond to the same object (false negative error). The second is that different objects may be contained within the same block (false-positive error). Considering the false-negative error, we use mse as our loss function. Thus, the network will only pull in features with the same index value (positive samples) and not push away features with different index values (negative samples). As shown in Table IX, a model pretrained using mse performs better in the downstream semantic segmentation task than one pretrained using infoNCE, which not only pulls in positive samples, but also pushes away negative samples. Moreover, we find that objects do not vary much within each scene in the SeCo dataset (see Fig. 10), implying that IndexNet is more prone to having the false-negative error. However, even so, IndexNet outperforms SeCo [46] in the downstream semantic segmentation task (see Table VII), which demonstrates that this design makes IndexNet not overly sensitive to the false-negative error and is therefore reasonable. For the false-positive error, we believe that using a ground-truth segmentation mask to guide the selection of positive samples is the best solution. However, this is not consistent with the intention of unsupervised pretraining. While it is possible to produce a segmentation mask using an unsupervised method, it is not computationally efficient and may still result in the same false-positive error. As a result, we believe that there is a specific correlation between adjacent pixels in RSIs. Thus, we argue that their corresponding features should also be closer in the feature space (e.g., roads and vehicles are often closer together, thus, it is reasonable to assume that if there are roads in the image, then vehicles are more likely to exist as well). With this perspective, we designed the block-wise index mask.

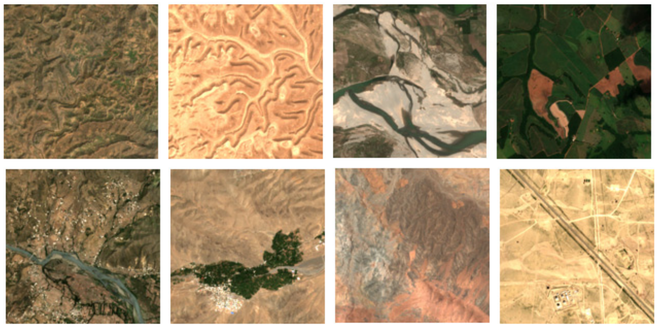

Fig. 10. **Sample images from the SeCo-100K dataset [46].**

Of course, we acknowledge that Index Contrast still has the potential for further improvement. For example, in our work, we used a fixed block size to align with the output feature map of ResNet, but better results may be obtained by adaptively determining the block size depending on the image's local properties. We leave this for further work.

B. Expectation

SSL has seen huge success in computer vision, but it has yet to be completely explored in the field of remote sensing. Since SSL is able to learn spatiotemporal invariance features without any annotated labels, it is ideal for applying it to RSIs, which vary greatly in time and location. Our method is just one possible implementation of SSL on RSIs. We believe that further refinements of the proposed index mask and algorithm that fully consider the characteristics of RSIs could lead to better performance.

VI. Conclusion

We introduced IndexNet, a novel SSL framework designed for RSI semantic segmentation task. Our method can capture both image-level and pixel-level spatiotemporal representations from large amounts of unlabelled data and perform better when fine-tuned to the semantic segmentation task. Moreover, the proposed method largely reduces the dependence on annotated data and bridges the gap between self-supervised pretraining and RSI semantic segmentation tasks. We expect the proposed method could be applied to large-scale remote sensing data to fully realize its potential, and our work could serve as a baseline for the pretraining method for semantic segmentation of RSIs.

References

[1] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 223–236, May 2019.

[2] M. Schmitt, J. Prexl, P. Ebel, L. Liebel, and X. X. Zhu, "Weakly supervised semantic segmentation of satellite images for land cover mapping—Challenges and opportunities," 2020, *arXiv:2002.08254*.

[3] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote Sens.*, vol. 11, no. 4, p. 403, 2019.

[4] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in SAR satellite images with deep fully convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1867–1871, Dec. 2018.

[5] A. Milioto, P. Lottes, and C. Stachniss, "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2229–2235.

[6] T. Anand, S. Sinha, M. Mandal, V. Chamola, and F. R. Yu, "AgriSegNet: Deep aerial semantic segmentation framework for IoT-assisted precision agriculture," *IEEE Sensors J.*, vol. 21, no. 16, pp. 17581–17590, Aug. 2021.

[7] K. Chen, K. Fu, M. Yan, X. Gao, X. Sun, and X. Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 173–177, Feb. 2018.

[8] R. Li *et al.*, "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.

[9] J. Liu, S. Wang, X. Hou, and W. Song, "A deep residual learning serial segmentation network for extracting buildings from remote sensing imagery," *Int. J. Remote Sens.*, vol. 41, no. 14, pp. 5573–5587, Jul. 2020.

[10] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, "ISPRS semantic labeling contest," ISPRS, Leopoldshöhe, Germany, 2014, vol. 1, p. 4.

[11] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[12] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, J. Vanschoren and S. Yeung, Eds., 2021, pp. 1–16. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf

[13] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 59–69, Apr. 2019.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[17] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[18] J.-B. Grill *et al.*, "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[19] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9912–9924.

[20] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3024–3033.

[21] O. J. Henaff, S. Koppula, J.-B. Alayrac, A. van den Oord, O. Vinyals, and J. Carreira, "Efficient visual pretraining with contrastive detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10086–10096.

[22] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16684–16693.

[23] S. Liu, Z. Li, and J. Sun, "Self-EMD: Self-supervised object detection without ImageNet," 2020, *arXiv:2011.13677*.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

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

[27] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, in Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 2015, pp. 1–14.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[30] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[31] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "OCNet: Object context for semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 2375–2398, Aug. 2021.

[32] H. Zhang, H. Zhang, C. Wang, and J. Xie, "Co-occurrent features in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 548–557.

[33] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS J. Photogramm. Remote Sens.*, vol. 170, pp. 15–28, Dec. 2020.

[34] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12416–12425.

[35] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 649–666.

[36] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[37] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 1096–1103.

[38] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 69–84.

[39] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=S1v4N2l0-

[40] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[41] S. Vincenzi *et al.*, "The color out of space: Learning self-supervised representations for earth observation imagery," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3034–3041.

[42] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1182–1191.

[43] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[44] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2598–2610, Mar. 2021.

[45] K. Ayush *et al.*, "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10181–10190.

[46] O. Manas, A. Lacoste, X. Giro-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9414–9423.

[47] P. Akiva, M. Purri, and M. Leotta, "Self-supervised material and texture representation learning for remote sensing tasks," 2021, *arXiv:2112.01715*.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[49] H. Li *et al.*, "Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images," 2021, *arXiv:2106.10605*.

[50] Z. Zheng, X. Zhang, P. Xiao, and Z. Li, "Integrating gate and attention modules for high-resolution image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4530–4546, 2021.

[51] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[52] V. G. T. da Costa, E. Fini, M. Nabi, N. Sebe, and E. Ricci, "Solo-learn: A library of self-supervised methods for visual representation learning," *J. Mach. Learn. Res.*, vol. 23, no. 56, pp. 1–6, 2022.

[53] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[54] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[55] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Dilxat Muhtar is currently pursuing the B.S. degree in geographic information science with Nanjing University, Nanjing, China.

His research interests include self-supervised and transfer learning for remote sensing.

Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information system and the Ph.D. degree in remote sensing of resources and the environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a Visited Student with the Informatics Institute, University of Missouri, Columbia, MO, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University, where he is currently an Associate Professor with the Department of Geographic Information Science. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.

Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018 and has been a Professor since 2019. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Hesse, Germany, from 2011 to 2012; and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. He has authored four books and over 160 articles. His research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.