DRSNet
Novel Architecture for Small Patch and Low-resolution Remote Sensing Image Scene Classification
Chen, Feihao; Tsou, Jin Yeu

Published in:
International Journal of Applied Earth Observation and Geoinformation

Published: 15/12/2021

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY-NC-ND

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1016/j.jag.2021.102577

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher’s definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
DRSNet: Novel architecture for small patch and low-resolution remote sensing image scene classification

Feihao Chen a, Jin Yeu Tsou a,b,*

a School of Architecture, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
b Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong Special Administrative Region

1. Introduction

Remote sensing (RS) image scene classification plays a remarkable role in various applications, such as land surface classification, urban planning, and risk management. For land use and cover classification, conventional approaches using spectral features of RS images have been widely adopted (Chen and Tsou, 2021). However, challenges arise concerning the accurate classification of highly heterogeneous land surface. Accordingly, unsatisfactory classification results could be obtained owing to interclass similarity and intraclass variation phenomena (Cheng et al., 2017a; Tong et al., 2018b).

In recent years, deep learning (DL) has drawn attention in the RS community because of its outstanding performance compared with conventional methods. DL is inspired by the biological brain and uses simple but massively interconnected units (i.e., neurons) to replicate cerebral mechanisms (Castelluccio et al., 2015). In addition to spectral and spatial features, DL has immense potential in exploiting semantic information from input images (Tong, 2018a). Convolutional neural network (CNN), which is the most well-known DL method for image recognition tasks, has been used in several published papers in the RS domain, with state-of-the-art results (Liu et al., 2019; Liu and Shi, 2020; Qiu et al., 2019; Qiu et al., 2020; Rosentreter et al., 2020; Yoo et al., 2019).

However, the majority of the said studies utilized very-high-resolution (VHR) satellite and aerial imagery for CNN training. For example, Tong et al. (2018b) proposed a high-spatial-resolution data set that consists of Gaofen-2 images (1 m/pixel) and used it for image recognition (Tong et al., 2018b). Yang et al. (2019) adopted manually labeled image patches with a resolution of 0.2 m/pixel to train and test their algorithms, and achieved satisfying land use and cover classification results (Yang et al., 2019). Mou et al. (2018) and Liu et al. (2019) utilized a public data set called Pavia University (1.3 m/pixel) for their classification tasks (Liu et al., 2019; Mou et al., 2017). According to a review research by (Ma et al., 2019), only a few studies have adopted medium- or low-resolution RS data for scene classification or image segmentation tasks. Although high-spatial-resolution images contain richer feature information than medium- and low-resolution products, the former is often not easily available and, in numerous cases, are cost
prohibitive for researchers, nongovernment organizations (NGOs), and governments of developing countries (Gram-Hansen et al., 2019). In real-world applications, freely accessible RS data are substantial to and preferred by these communities. Thus, the World Urban Database and Access Portal Tools (WUDAPT) project has been initiated. It is a community-based volunteer program aiming to develop a global database that captures information on urban form and function using free data sources (Bechtel et al., 2015). To date, land surface maps of more than 120 cities in different countries have been produced and uploaded to the WUDAPT portal after quality assessments (Ren et al., 2019). Nonprofit NGOs such as the United Nations Children’s Fund (UNICEF) have also encouraged scientists to conduct their research using freely available medium- or low-resolution RS data. In a project supported by the UNICEF, researchers utilized Sentinel-2 images and a classic CNN model to detect informal settlements (Gram-Hansen et al., 2019; Helber et al., 2018). Inspiring mapping results have been achieved under their cost-effective method, indicating the importance of their work.

In addition, the capacity of existing CNN structures for RS image recognition should be improved. These structures are mainly designed to intake daily RGB images, which are relatively different from RS images in terms of sourced sensors, spectral bands, and resolutions. Several researchers from the RS community prefer utilizing classic CNN models and fine-tuning them for their own tasks; however, studies have reported that these networks are outperformed by conventional descriptors in several RS classification tasks (Casteluccio et al., 2015; Nogueira et al., 2017). In a recent comparative study conducted by (Chen and Tsou, 2021), the random forest machine learning algorithm with a moving window outperformed CNN schemes (i.e., CaffeNet and GoogLeNet) in terms of land surface classification with low-resolution RS images. Moreover, lightweight CNN structure has been a trend in recent years (Li et al., 2018). Deep neural networks have evolved remarkably, but their computational complexity and resource consumption continue to increase (Cheng et al., 2018). Classic CNN architectures such as AlexNet and VGG are redundant regarding the total amount of network parameters, which poses a challenge in the deployment of such networks on memory-limited devices (Cheng et al., 2017b). Thus, a new CNN structure with a low-redundancy network design is encouraged. In summary, the potential of DL-based approaches has yet to be fully explored in the RS community, even though advanced CNN structures are continuously pushing the frontiers (Qu et al., 2020).

To address the said challenges, a novel CNN architecture called DRSNet for moderate- and low-resolution RS image scene classification is proposed in this work. This task is more challenging than ordinary RGB image classification considering the distinctive attributes of RS data and the complex spectral features of land surface objects. Several widely used open-sourced RS products include Landsat (30 m/pixel), Sentinel (10 m/pixel), CBERS-2 (20 m/pixel), and Envisat-ASAR (30 m/pixel). In this paper, free Landsat 8 images are downloaded from the US Geological Survey website (http://glovis.usgs.gov) to develop our training, validating, and testing data sets. The strengths of Inception-ResNet (Szegedy et al., 2017) are combined with the channel attention mechanism (Hu et al., 2018), and an effective block called residual inception channel attention (RICA) for feature extraction is proposed. Shortcut connections (He et al., 2016) are used to connect RICA blocks rather than convolutional layers densely. The reduction module (Szegedy et al., 2017) is also applied instead of normal pooling layers to prevent representational bottleneck (Szegedy et al., 2016) because our input data are modified with a small height and width. Upsampling layers are adopted before the final pooling layer to reduce information loss. To the best of our knowledge, this paper is the first to use such an upsampling structure for classification tasks. Network design tricks to reduce the parameters are utilized, including the use of special filters (e.g., 1 × 1, 1 × 3, and 3 × 1 filters) and reducing the number of linear layers. The proposed model with simple but innovative changes achieves better performance than state-of-the-art CNNs.

The main contributions of this paper are threefold.

(1) A new Landsat 8 data set for RS image recognition is constructed. Unlike the majority of the existing data sets, which consist of VHR satellite or aerial images, our data set features medium/low spatial resolution and small patch size. This data set provides the RS community with a new option to evaluate and advance various algorithms.

(2) A novel CNN architecture called DRSNet is proposed for RS image scene classification and achieves a higher classification accuracy than existing CNN models in our data set. The proposed network also outperforms its counterparts in several public RS data sets, demonstrating good generalization capacity.

(3) An effective, cost-efficient approach is delivered. Our model runs efficiently on a laptop and is monetarily effective because it uses freely available, easily accessible Landsat 8 imagery. This process may be pivotal for several individual researchers, NGOs, and governments of developing countries.

The remainder of this paper is structured as follows: Section 2 briefly reviews several milestone CNN structures and upsampling operations from which inspiration to design DRSNet is taken. Section 3 introduces our data sets and the proposed network as well as evaluation metrics and implementation details. Section 4 presents the experimental results and comparisons of the proposed network and other mainstream CNN models. Lastly, Section 5 provides the conclusions.

2. Related research

2.1. ResNet and DenseNet

As the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-2015 competition, ResNet revolutionized the CNN architecture race because it introduced the concept of residual learning, that is, shortcut connection (He et al., 2016). Theoretically, a deeper network generates better classification results. However, network degradation problem also occurs when increasing the network depth (He et al., 2016). Shortcut connections mitigate the degradation problem and reduce the computational complexity. The idea of shortcut connection is to connect the output of at least one convolutional layer to the original input (Kumar et al., 2020). Let $X_{i-1}$ and $X_i$ be the input and output of the $i$th convolutional layer $F_i$, respectively. Residual learning can be expressed using the following equation:

$$X_i = F_i (X_{i-1}) + X_{i-1}$$  \hspace{1cm} (1)

Inspired by ResNet, DenseNet ensures maximum information paths between layers by connecting all layers (Jang and Park, 2019). Each layer in DenseNet takes all preceding feature maps as input and skip-connects the feature maps of each layer for all subsequent layers (Qu et al., 2020). The output $X_i$ of the $i$th convolutional layer $F_i$ can be denoted as follows:

$$X_i = F_i ([X_{0}, X_1, \ldots, X_{i-1}])$$  \hspace{1cm} (2)

where $[\cdot]$ represents feature concatenation along the channel axis.

DenseNet is effective in solving the vanishing gradient problem owing to its densely connected feature maps (Huang et al., 2017). ResNet and DenseNet have advantages in back-propagation with strong gradients, and DenseNet is often considered an extreme version of ResNet. However, the two networks have substantial differences. On the one hand, ResNet applies shortcut connection and refines the feature value by element-wise addition. On the other hand, DenseNet adopts dense skip connection and memorizes the feature value by channel concatenation.

2.2. Inception-ResNet

GoogLeNet, also known as Inception-v1, is the winner of the ILSVRC-
2014 competition. It shows that in addition to depth, the width of a network is a key factor of model performance. The essential component of GoogLeNet is its inception block (Szegedy et al., 2015). An inception block increases the width of a network and extracts features at different scales because it encapsulates several various-size kernels (i.e., $1 \times 1$, $3 \times 3$, and $5 \times 5$) in one module. It also reduces the number of learnable parameters by applying $1 \times 1$ kernels. As the newest version of the Inception-series networks, Inception-ResNet (Szegedy et al., 2017) combines the advantages of Inception architecture and ResNet, thereby improving the performances of both. Filter concatenation of the Inception block is replaced by residual connection. Experimental results indicate that Inception-ResNet converges more rapidly than Inception architectures.

2.3. Channel attention mechanism

Attention mechanism is proposed to mimic the human perception system. Attention modules of CNNs increase representation power by focusing on important features and suppressing unnecessary ones (Woo et al., 2018). They integrate global contextual dependencies in the spatial and channel dimensions (Qiu et al., 2020). Similar to the squeeze and excitation block of SENet (Hu et al., 2018), the channel attention system. Attention modules of CNNs increase representation power by multiple input activations within a filter window to a single activation, transposed convolutional layers associate a single input activation with multiple outputs (Badrinarayanan et al., 2017; Noh et al., 2015). Transposed convolution can retrieve the lost information introduced by convolution, pooling, and other downsampling processes (Mohammadimanesh et al., 2019). However, this process only reconstructs the size of feature maps and not the contained values. The retrieved feature maps are different from the original ones, although they share the same height and width (see Fig. 3).

The outputs of unsampling and unpooling operations are enlarged but sparse activation maps (Li et al., 2019; Mou et al., 2017). Unlike unsampling and unpooling, transposed convolution generates enlarged, dense feature maps. In addition, filters of transposed convolutional operations are learnable, which indicates that nonlinear upsampling results can be acquired (Long et al., 2015). Theoretically, transposed convolution with proper parameter settings and adequate trainings produces more precise reconstructed feature maps than unsampling and unpooling. Therefore, transposed convolution for upsampling operations is adopted here.

3. Proposed method

Our model is trained and tested on four data sets, namely, our own Landsat 8, EuroSAT (Helber et al., 2019), Brazilian Coffee Scenes, and UCMerced Land Use (Yang and Newsam, 2010). The last three public data sets are used to investigate the generalization capacity of DRSNet. Several state-of-the-art CNNs are selected as baseline models and tested on these data sets. The classification results of all networks will be compared and discussed in Section 4.

3.1. Data sets and preprocessing

3.1.1. Landsat 8 data set

A data set consisting of low-resolution, small-size Landsat 8 images, which cover the entire area of Dongguan City, China, is developed (Chen and Tsou, 2021). The local climate zone (LCZ) classification system (Stewart and Oke, 2012) is adopted in the current paper. The proposed data set contains 13 out of 17 subclasses of the LCZ category system, namely, LCZ 2–6, LCZ8, LCZ10, LCZ A–B, and LCZ D–G. Our data set has seven urban categories (i.e., LCZ 2–6, LCZ 8, and LCZ 10) and six natural categories (i.e., LCZ A–B, LCZ D–G).

Landsat 8 is an RS product with a spatial resolution of 30 m/grid. To ensure the minimum input size for CNNs, the origin Landsat image is resampled from 30 m/grid to 10 m/grid. The RS image is clipped into

![Fig. 1. Illustration of max-pooling (left) and unsampling (tight). Pooling is performed with a kernel of 2 × 2 and a stride of 2.](image-url)
small patches that are labeled carefully. The height and the width of each sample are 10. A total of 2432 and 809 patches are labeled for model training and testing, respectively, and are triple checked by senior professionals in the RS domain to ensure the quality of our data set.

Given the relatively low spatial resolution and the small input size of our RS data, our priority is to decrease the chance of overfitting. Data augmentation is applied to increase the total number of training samples. Each sample is rotated 90°, 180°, and 270°. Then, the four obtained samples are flipped. Therefore, the number of training samples is enlarged eight times (see Fig. 4), that is, the total number of training patches is 19456. About 20% of the training samples are randomly selected as the validation set to adjust the hyperparameters. L2 regularization and dropout are also adopted (i.e., weight decay and dropout rate are set as 0.0005 and 0.2, respectively). Input images are resized to 32 × 32 before feeding into CNNs.

Table 1 shows the detailed information of the data set, including the descriptions of each class, total number of training and testing samples, and number of samples of each class.

3.1.2. EuroSAT, Brazilian Coffee Scenes, and UCMerced land use data set

Proposed by (Helber et al., 2019), EuroSAT is developed for large-scale land use and cover classification. It is a data set based on Sentinel-2 satellite images, consisting of 27,000 labeled images with 10 different categories. The 10 categories are annual crop, forest, herbaceous vegetation, highway, industrial buildings, pasture, permanent crop, residential buildings, river, and sea and lake. Each class contains 200–300 images. Similar to our Landsat 8, EuroSAT has a low resolution (10 m/pixel) and a relatively small image size (64 × 64), making it an ideal data set to test the generalization capacity of DRSNet. EuroSAT is randomly divided into two parts with an 80/20 training–testing rate. About 20% samples are also randomly selected from the training set as the validating set to adjust the hyperparameters. DRSNet and other baseline CNNs are trained on EuroSAT from scratch. To accelerate the training, input images are resized from 64 × 64 to 32 × 32 for CNNs.

Brazilian Coffee Scenes, which is composed of scenes taken by the SPOT sensor (10 m/pixel) in 2005 over four counties in the State of Minas Gerais, Brazil, namely, Arceburgo, Guaranésia, Guaxupé, and Monte Santo, is proposed by (Penatti et al., 2015). It is a very challenging data set because of many intraclass variances caused by different...
crop management techniques (Penatti et al., 2015). Brazilian Coffee Scenes contains 2876 images with 64 × 64 pixels equally divided into two classes (coffee and noncoffee). The images in this data set are not optical (green–red–infrared instead of red–green–blue). An 80/10/10 training–validating–testing rate is adopted to train and test DRSNet and other baseline CNNs. Images of this data set are not optical (green–red–infrared instead of red–green–blue). An 80/10/10 training–validating–testing rate is adopted to train and test DRSNet and other baseline CNNs. Images of this data set are resized to 32 × 32 for model training, validating, and testing.

Different from Landsat 8, EuroSAT, and Brazilian Coffee Scenes, UCmerced Land Use consists of 2100 high-resolution (1 ft/pixel), large-size (256 × 256) aerial images (Yang and Newsam, 2010). Each of the following 21 classes have 100 images: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. UCmerced Land Use is utilized to test the performance of DRSNet on large-size, high-resolution RS images. This data set is divided into two parts following an 80/20 training–testing rate (i.e., for each class, 80 and 20 images are used for training and testing, respectively). The training set is enlarged eight times using data augmentation techniques. About 20% of the training samples are randomly selected to adjust the hyperparameters. DRSNet and other CNNs are also trained on this data set from scratch. Images of this data set are resized to 64 × 64

<table>
<thead>
<tr>
<th>Categories</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>LCZ 2: Compact midrise</td>
<td>245</td>
</tr>
<tr>
<td>LCZ 3: Compact low-rise</td>
<td>170</td>
</tr>
<tr>
<td>LCZ 4: Open high-rise</td>
<td>161</td>
</tr>
<tr>
<td>LCZ 5: Open midrise</td>
<td>116</td>
</tr>
<tr>
<td>LCZ 6: Open low-rise</td>
<td>181</td>
</tr>
<tr>
<td>LCZ 8: Large low-rise</td>
<td>267</td>
</tr>
<tr>
<td>LCZ 10: Heavy industry</td>
<td>153</td>
</tr>
<tr>
<td>LCZ A: Dense trees</td>
<td>264</td>
</tr>
<tr>
<td>LCZ B: Scattered trees</td>
<td>150</td>
</tr>
<tr>
<td>LCZ D: Low plants</td>
<td>149</td>
</tr>
<tr>
<td>LCZ E: Bare rock or paved</td>
<td>107</td>
</tr>
<tr>
<td>LCZ F: Bare soil or sand</td>
<td>132</td>
</tr>
<tr>
<td>LCZ G: Water</td>
<td>337</td>
</tr>
<tr>
<td>Total</td>
<td>2432</td>
</tr>
</tbody>
</table>

Table 1: Details of the proposed Landsat 8 data set. Each sample represents a 10 m × 10 m region.
as network inputs.

3.2. DRSNet architecture

To learn the features of training samples effectively, a novel CNN architecture that consists of a stem module, several feature-learning modules, reduction modules, upsampling steps, and a final global pooling layer is proposed.

3.2.1. Stem module

The stem module serves as a low-level feature extractor, abstracting initial input information for the following layers. Given that our data set is featured with a low resolution and a small patch size, the stem module uses several convolutional branches with different-size kernels to capture information from various scales (Fig. 5). For each branch, a $1 \times 1$ filter is used to adjust the number of output channels, reducing the redundancy of the network parameters. Special kernel forms (i.e., $1 \times 3$, $3 \times 1$, $1 \times 5$, and $5 \times 1$) are applied to increase the receptive fields while maintaining a relatively few trainable parameters. Research has emphasized the importance of the depth (Simonyan and Zisserman, 2014) and the width (Szegedy et al., 2015; Zagoruyko and Komodakis, 2016) of a model for its performance. Thus, two parallel convolutional branches (i.e., $1 \times 3$ and $3 \times 1$) are designed and concatenated to increase the width of the stem module. Two $1 \times 5$ and $5 \times 1$ convolutional layers are stacked to increase the depth of the module. All these branches are eventually concatenated to extract the input features fully.

3.2.2. Feature-learning Module: RICA

Inspired by Inception-ResNet and the channel-attention mechanism, this paper proposes a novel module called RICA to learn high-level, intrinsic features. Two different forms of the RICA module (i.e., RICA I and RICA II) are designed. Fig. 6a and 6b show the structures of the two forms. The RICA I block includes three different convolutional branches, which is similar to the structure of an inception module. RICA II contains two convolutional branches and utilizes large kernels (i.e., $1 \times 7$ and $7 \times 1$) to include additional global information, considering that the sizes of the feature maps are small. The outputs of these branches are concatenated (as illustrated in Fig. 6a and 6b, which is followed by a $1 \times 1$ convolution layer to increase the nonlinearity of the module and reduce the number of output channels. A squeeze and excitation step is performed to exploit the potentiality of input channels. First, a global pooling operation is applied to obtain the channel-wise statistics. Second, two $1 \times 1$ convolutional layers with channel reduction ratio $r$ are performed to increase the nonlinearity of the module and reduce the overall complexity. Third, a gating mechanism with sigmoid function is used to obtain the weighted statistics. The original feature maps are multiplied by the statistics to obtain the channel-wise weighted feature maps. Lastly, element-wise addition between the original and weighted feature maps is conducted to obtain the final feature maps.

3.2.3. Reduction module

Downsampling steps, such as pooling layers, are effective in terms of increasing the receptive field and reducing computational complexity. However, they will inevitably lead to information loss, particularly for samples with small heights and widths. Several studies have proposed to use two different pooling layers (e.g., max-pooling and average-pooling) simultaneously and concatenate their outputs to mitigate this effect (Qiu et al., 2020). The current paper adopts the idea of (Szegedy et al., 2017) and proposes a reduction module consisting of a pooling layer and several convolutional branches (Fig. 7). They are concatenated to ensure that sufficient information will be passed to the next layer.

3.2.4. Upsampling layers

Before the final global average pooling layer, a simple but innovative step (i.e., upsampling layers) is added into our network. Given that the training and testing samples are featured with coarse resolution and small size, valuable information may be lost after several feature-extraction blocks and reduction modules. Fig. 8 shows that two consecutive $4 \times 4$ transposed convolutional layers are utilized to retrieve lost information. Each transposed convolutional layer is followed by a $1 \times 1$ convolutional layer. These $1 \times 1$ convolutional layers are used to reduce network parameters and increase model nonlinearity. The outputs of the $1 \times 1$ convolutional layers and previous feature maps are connected by performing element-wise addition, thereby enriching the information for final global pooling. To the best of our knowledge, this paper is the first to utilize an upsampling structure for image scene classification tasks. The advantages of such a CNN structure will be tested on several data sets.

Fig. 9 shows the overall structure of our proposed network. Inspiration is taken from DenseNet, and dense shortcut connections are applied to prevent network degradation. However, our network and DenseNet have two major differences regarding dense connection. DenseNet densely connects convolutional layers, whereas the proposed network densely connects RICA modules rather than convolutional

---

Fig. 5. Architecture of stem module. Special zero paddings are utilized to keep the sizes of feature maps unchanged. Each convolutional layer is followed by a BN layer and ReLU. H and W represent the height and the width of the input feature maps, respectively. The size of the feature maps remains unchanged during the whole module.
layers. In addition, DenseNet connects layers by channel concatenation, whereas our model connects modules via element-wise addition.

3.3. Baseline CNNs

A variety of CNNs are selected as baselines, including AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), Inception-ResNet-v2 (Szegedy et al., 2017), SENet (Hu et al., 2018), and EfficientNet (Tan and Le, 2019). AlexNet is the winner of the ILSVRC-2012 competition, and its success has accelerated the research on the structure of CNN models (Khan et al., 2020). VGG is composed of convolutional layers with $3 \times 3$ filters, $2 \times 2$ max-pooling layers with a stride of 2, and three fully connected layers at the end (Qiu et al., 2020). ResNet improves the training efficiency of very deep neural networks by introducing shortcut connections (Qiu et al., 2020). As the latest version of Inception-ResNet, Inception-ResNet-v2 achieves the highest classification accuracies on ImageNet data set. SENet is the champion of the last ILSVRC competition, and its core part is the channel-attention module. The compound scaling method is adopted to develop EfficientNet (Tan and Le, 2019), and this network exhibits much better accuracy and efficiency than previous CNNs. These baselines are selected to have a varying number of layers and parameters, to represent a wide range of cases.

Fig. 6. a. Architecture of RICA I module. $C$ is the number of channels, and $r$ is the channel reduction ratio. In this paper, $C = 448$, and $r$ is set to 16. $H$ and $W$ are the height and the width of the input feature maps, respectively. The size of the feature maps remains unchanged during the whole module. Fig. 6b. Architecture of RICA II module. Compared with RICA I, RICA II has two convolutional branches. RICA II adopts large kernels to capture additional global information. The size of the feature maps remains unchanged during the whole module.
3.4. Evaluation metrics

Metrics used for performance assessments include cross-entropy loss, overall accuracy (OA), and Kappa coefficient. The cross-entropy loss function quantifies how well the model outputs match the ground truth labels in the training/testing set. This metric is often used to determine whether the model is fully trained and when to stop the training. It can be expressed using the following equation:

\[ H(y, p) = - \sum_i y_i \log(p_i) \]  

where \( y \) is the one-hot encoded vector of the labels, \( i \) represents a certain class, and \( p \) is the predicted probability distribution, which represents the likelihood of belonging to a certain class. Moreover, \( p \) can be...
formulated as follows:
\[ p_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \]  
where \( z \) is the C-dimensional vector of the real-valued scores of the output layer of the model, and \( C \) is the total number of classes.

OA is the percentage of correctly classified patches and all patches in the entire testing data set, and Kappa measures the agreement between the prediction and ground truth in the testing set. High OA and kappa indicate a satisfactory classification result. \( P_{ab} \) represents the number of samples of class \( a \) that are predicted to belong to class \( b \), \( t_a = \sum P_{ab} \) computes the total number of patches that belong to class \( a \), \( c \) is an integer that belongs to \([1, C]\), and \( C \) is the number of categories.

\[ OA = \frac{\sum a P_{aa}}{\sum a t_a} \]  

**Fig. 10.** Schematic process flow of this paper.
\[
Kappa = \frac{P_o - P_e}{1 - P_e} \tag{7}
\]

where \[
P_o = \sum \frac{P_{aa}}{t_a} \tag{8}
\]
\[
P_e = \frac{\sum (\sum b P_{cb} \times \sum a P_{ac})}{\sum t_a \times \sum t_a} \tag{9}
\]

In addition, the total number of parameters (Params), parameter size (Param_size), and floating point operations (Flops) of each network are compared. Params and Param_size are often used to measure the consumption of computer memory by a CNN model. A larger Params/Param_size indicates more consumption of computer memory resources. The forward propagation of a neural network is multiplication and accumulation; thus, Flops refers to the number of multiplication and addition operations in forward reasoning. It is often used to evaluate the computational complexity of a model or an algorithm.

### 3.5. Implementation details

For DRSNet, a combination of [RICA II, RICA II, RICA I, RICA II] is used as the feature-learning modules (see Fig. 9). Filter weights are initialized using the Kaiming initialization (He et al., 2015). The kernel sizes of the convolutional layers are set to \(3 \times 3\), with a few special kernels (e.g., \(1 \times 7, 7 \times 1\)) as exceptions. Given that the heights and widths of the input patches are small, only the sizes of the feature maps in the reduction modules are reduced. Zero-padding is adopted to keep their sizes unchanged in the stem and feature-learning modules. Each convolutional layer is followed by a batch normalization layer and ReLU.

The proposed DRSNet and baseline CNNs with the PyTorch framework are implemented, and an NVIDIA GEFORCE RTX 2060 GPU is used to accelerate the training. Training DRSNet on our data set takes several hours because the total amount of network parameters is relatively small. The training and testing batch sizes are 54. The initial learning rate is 0.008, and a stochastic gradient descent optimizer with a weight decay of 0.0005 and a momentum of 0.9 is applied. The initial learning rate is divided by 10 after every 10 epochs. Each network is trained 50 epochs on one data set. After the training, the test accuracy from the saved weights with the highest validation accuracy for each network is reported.

The training configurations (i.e., initial learning rate, total epochs, batch size, and dropout rate) of baseline networks are the same as DRSNet except for Inception-ResNet-v2. Owing to the limited computational memory of our laptop and the complexity of the network structure, the batch size of Inception-ReNet-v2 has to be reduced to 26 (batch sizes of other networks are 54). Fig. 10 shows the major steps of this paper.

### 4. Experimental results and discussion

Comparative experiments are conducted to investigate the effectiveness and the generalization ability of DRSNet. Experimental results are given in this section.

#### 4.1. Landsat 8 data set

DRSNet achieves the best results on our Landsat 8 data set. Figs. 11 and 12 show the testing loss and the accuracy of the proposed DRSNet, respectively. Fig. 11 illustrates that testing loss decreases rapidly before the first 20 epochs, and steadily goes down afterward. It fluctuates slightly in the last 15 epochs, indicating that the network has been fully trained. Correspondingly, OA increases rapidly after a temporary decrease and reaches a plateau afterward. It slightly fluctuates at
Fig. 13 presents the confusion matrix of DRSNet. In summary, DRSNet can accurately classify the proposed Landsat 8 data set, with the accuracies of the eight categories exceeding or nearly 90%. This classification result is impressive considering the considerable interclass similarity and intraclass variation phenomenon as well as the low resolution of the input images. DRSNet performs extremely well in all the natural categories except for LCZ E (70.00%). This result can be partially explained by the relatively few training and testing samples of LCZ E. In terms of urban categories, DRSNet achieves satisfactory results in LCZ 8 (93.59%), LCZ 2 (93.22%), and LCZ 4 (89.36%). The accuracies of LCZ 6 (79.25%), LCZ 3 (76.47%), LCZ 10 (71.43%), and LCZ 5 (70.00%) are acceptable because even experts in this domain could be easily confused with these categories. Fig. 13 shows that four out of the 34 testing patches of LCZ 3 (compact low-rise) are misclassified into LCZ 2 (compact midrise), and six out of the 30 patches of LCZ 5 (open midrise) are wrongly classified into LCZ 6 (open low-rise).

Fig. 14 shows the losses of all CNN schemes. As can be seen, the losses of SENet and ResNet34 increase after approximately 15 and 20 epochs, respectively, indicating that the two networks exhibit overfitting. The loss curve of Inception-ResNet-v2 presents a high variance, and its final value is approximately 12, which is higher than that of other CNNs. EfficientNet-b0 has a faster convergence speed than EfficientNet-b7, and their final losses are slightly lower than that of ResNet34 and ResNet50. VGG19 has a convergence speed similar to EfficientNet-b0, and its final loss is considerably smaller than that of AlexNet. Compared with existing CNNs, DRSNet has a rapid convergence speed and achieves the lowest final testing loss.

Table 2 displays the OA, Kappa, and accuracies of each category of all networks. DRSNet outperforms other CNNs and achieves the highest OA (90.36%) and Kappa (0.8902). VGG19 is the runner-up, with OA of 88.5% and Kappa of 0.8690. EfficientNet-b7 is third, followed by SENet. The performances of ResNet34 and ResNet50 are not outstanding and are outperformed by DRSNet, VGG19, EfficientNet-b7, and SENet. Inception-ResNet-v2 achieves OA of approximately 5% lower than DRSNet and 3% lower than VGG19.

Table 2 also shows that DRSNet achieves the highest accuracies of several subclasses, namely, LCZ 2, LCZ 4, LCZ B, LCZ E, LCZ F, and LCZ G. In such categories as LCZ 8 and LCZ D, DRSNet achieves the second highest classification results. Such networks as SENet and VGG also perform well in several classes (i.e., LCZ 8 and LCZ A). The accuracies of LCZ 10 and LCZ E are unsatisfactory for all CNNs because the two categories could be easily misclassified.

4.2. EuroSAT data set

The classification accuracies of Inception-ResNet-v2, DRSNet, and SENet on EuroSAT are extremely high, and these networks outperform other CNNs by a large margin. Among them, OA (97.07%) and Kappa (0.9674) of Inception-ResNet-v2 are the highest, followed by DRSNet (96.91% and 0.9629, respectively) and SENet (96.50% and 0.9610, respectively). EfficientNet-b0 ranks fourth, with OA of 88.85% and Kappa of 0.8759. The remainder of the networks achieve OAs from approximately 81% to 87% and Kappa from approximately 0.8 to 0.85. Table 3 shows the detailed classification results of these networks.

Although OA and Kappa of DRSNet are slightly outperformed by Inception-ResNet-v2, it exhibits strong recognition capacities. Table 3 shows that Inception-ResNet-v2 leads in six out of 10 categories. SENet outperforms other networks in one category (i.e., highway). DRSNet achieves the highest classification accuracies of four categories, namely, annual crop (97.67%), industrial buildings (97.20%), permanent crop (92.80%), and river (96.80%). The accuracies of each class of DRSNet are over 92%, indicating its good feature-learning abilities for all classes in EuroSAT.

4.3. Brazilian Coffee Scenes data set

DRSNet also achieves satisfactory classification results on Brazilian Coffee Scenes, a data set consisting of nonoptical images with relatively low spatial resolution and small image size. Table 4 shows that the OA (91.67%) and Kappa (0.8333) of DRSNet are the highest among all CNNs. SENet is a close follower, with OA of 90.97% and Kappa of 0.8194. EfficientNet-b0 and ResNet34 exhibit similar classification performance regarding OA (89.93% and 89.24%, respectively) and Kappa (0.7986 and 0.7847, respectively). A complex CNN does not necessarily achieve high accuracies on this data set, and Inception-ResNet-v2 and EfficientNet-b7 are examples. The Brazilian Coffee Scenes data set only has two categories, namely, coffee and noncoffee. SENet and AlexNet achieve the highest accuracies on coffee and noncoffee, respectively. DRSNet achieves the second highest accuracies on
exhibits strong feature-learning abilities, achieving highest accuracies in classification accuracy of 100.00% in several categories. SENet also outperforms DRSNet by approximately 2.6% for OA and 3% for Kappa. AlexNet, VGG, ResNet34 ranks third (i.e., OA is 74.05%, and Kappa is 0.7275). OAs of baseline CNNs (Inception-ResNet-v2, SENet, EfficientNet-b0, EfficientNet-b7) are approximately 60% to 70% (i.e., 59.05% [ResNet50], 62.62% [EfficientNet-b0], and 70.00% [Inception-ResNet-v2]). AlexNet, VGG, and EfficientNet-b7 achieve poor OAs (i.e., 38.81%, 18.33%, and 43.10%, respectively). Table 3 presents the detailed results.

Table 3: Classification results of DRSNet and baseline CNNs on the Brazilian Coffee Scenes data set.

<table>
<thead>
<tr>
<th></th>
<th>DRSNet</th>
<th>AlexNet</th>
<th>VGG19</th>
<th>ResNet34</th>
<th>ResNet50</th>
<th>Inception-ResNet-v2</th>
<th>SENet</th>
<th>EfficientNet-b0</th>
<th>EfficientNet-b7</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA</td>
<td>90.36%</td>
<td>81.71%</td>
<td>88.50%</td>
<td>87.39%</td>
<td>87.02%</td>
<td>85.78%</td>
<td>87.89%</td>
<td>87.02%</td>
<td>88.01%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.8902</td>
<td>0.7913</td>
<td>0.8690</td>
<td>0.8565</td>
<td>0.8523</td>
<td>0.8384</td>
<td>0.8621</td>
<td>0.8522</td>
<td>0.8634</td>
</tr>
<tr>
<td>LCZ 2</td>
<td>93.22%</td>
<td>74.58%</td>
<td>83.05%</td>
<td>61.02%</td>
<td>71.19%</td>
<td>71.19%</td>
<td>72.88%</td>
<td>69.49%</td>
<td>79.66%</td>
</tr>
<tr>
<td>LCZ 3</td>
<td>76.47%</td>
<td>32.35%</td>
<td>70.59%</td>
<td>85.29%</td>
<td>79.41%</td>
<td>85.29%</td>
<td>82.35%</td>
<td>76.47%</td>
<td>61.76%</td>
</tr>
<tr>
<td>LCZ 4</td>
<td>89.36%</td>
<td>82.98%</td>
<td>82.98%</td>
<td>87.23%</td>
<td>76.60%</td>
<td>80.85%</td>
<td>78.72%</td>
<td>78.72%</td>
<td>74.47%</td>
</tr>
<tr>
<td>LCZ 5</td>
<td>70.00%</td>
<td>36.67%</td>
<td>83.33%</td>
<td>80.00%</td>
<td>66.67%</td>
<td>73.33%</td>
<td>83.33%</td>
<td>73.33%</td>
<td>76.67%</td>
</tr>
<tr>
<td>LCZ 6</td>
<td>79.25%</td>
<td>69.81%</td>
<td>75.47%</td>
<td>75.47%</td>
<td>69.81%</td>
<td>52.83%</td>
<td>66.04%</td>
<td>69.81%</td>
<td>84.91%</td>
</tr>
<tr>
<td>LCZ 8</td>
<td>93.55%</td>
<td>94.87%</td>
<td>94.87%</td>
<td>92.31%</td>
<td>94.87%</td>
<td>88.46%</td>
<td>92.31%</td>
<td>91.03%</td>
<td>94.87%</td>
</tr>
<tr>
<td>LCZ10</td>
<td>71.43%</td>
<td>47.62%</td>
<td>78.57%</td>
<td>66.67%</td>
<td>71.43%</td>
<td>71.43%</td>
<td>69.05%</td>
<td>71.43%</td>
<td>78.57%</td>
</tr>
<tr>
<td>LCZ A</td>
<td>93.75%</td>
<td>93.75%</td>
<td>93.75%</td>
<td>95.83%</td>
<td>92.71%</td>
<td>92.71%</td>
<td>95.83%</td>
<td>94.79%</td>
<td>93.75%</td>
</tr>
<tr>
<td>LCZ B</td>
<td>94.29%</td>
<td>94.13%</td>
<td>94.13%</td>
<td>88.57%</td>
<td>94.29%</td>
<td>88.57%</td>
<td>94.29%</td>
<td>91.43%</td>
<td>85.71%</td>
</tr>
<tr>
<td>LCZ D</td>
<td>93.18%</td>
<td>81.82%</td>
<td>93.18%</td>
<td>90.91%</td>
<td>95.45%</td>
<td>86.64%</td>
<td>86.64%</td>
<td>93.18%</td>
<td>88.64%</td>
</tr>
<tr>
<td>LCZ E</td>
<td>70.00%</td>
<td>45.00%</td>
<td>70.00%</td>
<td>65.00%</td>
<td>70.00%</td>
<td>70.00%</td>
<td>70.00%</td>
<td>65.00%</td>
<td>65.00%</td>
</tr>
<tr>
<td>LCZ F</td>
<td>87.72%</td>
<td>77.19%</td>
<td>71.93%</td>
<td>82.46%</td>
<td>80.70%</td>
<td>85.96%</td>
<td>87.72%</td>
<td>85.96%</td>
<td>84.21%</td>
</tr>
<tr>
<td>LCZ G</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

4.4. UCMerced land use data set

Although DRSNet is designed for low-resolution, small-size RS imagery classification tasks, it exhibits impressive results on UCMerced Land Use, indicating its strong generalization capacity for high-spatial-resolution data sets. The OA and Kappa of DRSNet (i.e., 83.33% and 0.8250, respectively) are the highest compared with the baseline CNNs. SENet follows with OA of 80.71% and Kappa of 0.7975. DRSNet outperforms SENet by approximately 2.6% for OA and 3% for Kappa. ResNet34 ranks third (i.e., OA is 74.05%, and Kappa is 0.7275). OAs of ResNet50, EfficientNet-b0, and Inception-ResNet-v2 range from approximately 60% to 70% (i.e., 59.05% [ResNet50], 62.62% [EfficientNet-b0], and 70.00% [Inception-ResNet-v2]). AlexNet, VGG, and EfficientNet-b7 achieve poor OAs (i.e., 38.81%, 18.33%, and 43.10%, respectively). Table 4 presents the detailed results.

Table 4: Classification results of DRSNet and baseline CNNs on the UCMerced land use data set.

<table>
<thead>
<tr>
<th></th>
<th>DRSNet</th>
<th>AlexNet</th>
<th>VGG19</th>
<th>ResNet34</th>
<th>ResNet50</th>
<th>Inception-ResNet-v2</th>
<th>SENet</th>
<th>EfficientNet-b0</th>
<th>EfficientNet-b7</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA</td>
<td>91.67%</td>
<td>86.11%</td>
<td>87.85%</td>
<td>89.24%</td>
<td>86.46%</td>
<td>88.89%</td>
<td>90.97%</td>
<td>89.93%</td>
<td>86.11%</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.8333</td>
<td>0.7222</td>
<td>0.7569</td>
<td>0.7847</td>
<td>0.7292</td>
<td>0.7778</td>
<td>0.8194</td>
<td>0.7986</td>
<td>0.7222</td>
</tr>
<tr>
<td>Coffee</td>
<td>91.67%</td>
<td>79.86%</td>
<td>88.19%</td>
<td>88.89%</td>
<td>88.19%</td>
<td>88.89%</td>
<td>92.36%</td>
<td>90.28%</td>
<td>90.28%</td>
</tr>
<tr>
<td>Noncoffee</td>
<td>91.67%</td>
<td>92.36%</td>
<td>87.50%</td>
<td>89.58%</td>
<td>84.72%</td>
<td>88.89%</td>
<td>89.58%</td>
<td>89.58%</td>
<td>81.94%</td>
</tr>
</tbody>
</table>

both classes.

4.5. Discussion

In this paper, DRSNet and baseline CNNs are trained and tested on the proposed Landsat 8 and three other public RS data sets (i.e., EuroSAT, Brazilian Coffee Scenes, and UCMerced). DRSNet achieves the best classification results on Landsat 8, Brazilian Coffee Scenes, and UCMerced, and the second highest OA and Kappa on EuroSAT. The outstanding performance of DRSNet on the Landsat 8 data set indicates its strong abilities for RS scene classification tasks. The OA of DRSNet is 2%–9% higher than that of baseline CNNs, and its Kappa is 2%–10% higher than that of baseline CNNs. Such networks as Inception-ResNet-v2, SENet, and EfficientNet report state-of-the-art classification results on ImageNet, but they are outperformed by DRSNet on the proposed Landsat 8. A reasonable explanation is that our data set features a low resolution and a small image size, which is relatively different from ImageNet. Normal convolutional layers with large kernels and strides lead to losing excessive information during the feature-extraction steps. Unlike baseline CNNs, DRSNet applies well-designed modules to address these challenges. All modules of DRSNet follow the philosophy of utilizing small kernels and impressive results have been reported. Therefore, a CNN architecture suitable for RS recognition tasks is remarkable for researchers in this domain because existing CNNs may not perform well in these tasks.
is not as competitive as DRSNet in terms of classification accuracy and number of parameters and parameter sizes of DRSNet and baseline CNNs. and EfficientNet-b0 are smaller than DRSNet, EfficientNet-b0 also an effective way of controlling parameter size. Although Param_size works except for EfficientNet-b0. Our philosophy of using small kernels reduces Params and Param_size. Reducing the number of linear layers is an effective way of controlling parameter size, and exploiting the potentiality of different channels. Upsampling operations between feature-learning modules enrich information for final global pooling.

Experimental results also reveal the good generalization capacities of DRSNet. DRSNet exhibits impressive classification accuracies on EuroSAT and Brazilian Coffee Scenes. These two data sets are similar to our Landsat eight data sets in terms of spatial resolution and patch size. Particularly, Brazilian Coffee Scenes is a special, challenging data set. Images of this data set are not optical, and the intraclass variation phenomenon is severe. In addition, DRSNet outperforms all baseline CNNs on UCMerced, a public data set of high-resolution, large-size RS images, showing its reliable classification capacity. In summary, DRSNet is a satisfactory architecture for small patch size RS scene classification tasks. DRSNet provides researchers in the RS community with an alternative because existing state-of-the-art CNNs may not perform well in several RS-related tasks.

The parameter size and total number of trainable parameters of DRSNet are considerably less than those of the baseline CNNs. Table 6 presents the two metrics of all CNNs and shows that the Params of DRSNet is 12.03 million, which is only half of that of ResNet50. The Param_size of DRSNet is 45.88 MB, a value lower than those of all networks except for EfficientNet-b0. Our philosophy of using small kernels reduces Params and Param_size. Reducing the number of linear layers is also an effective way of controlling parameter size. Although Param_size and Params of EfficientNet-b0 are smaller than DRSNet, EfficientNet-b0 is not as competitive as DRSNet in terms of classification accuracy and generalization ability. Compared with the baseline CNNs, DRSNet exhibits potentiality for deployment in devices with low memory resources.

Table 6 also displays the Flops of DRSNet and other baseline CNNs. The Flops of Inception-ResNet-v2 is the highest among all CNN schemes (i.e., $7.2 \times 10^8$), indicating that it has the highest model complexity. It is followed by DRSNet, which requires approximately $5.0 \times 10^8$ Flops in a single forward pass for a $32 \times 32$ pixel input image. SENet has Flops of $1.2 \times 10^8$. The Flops of these three networks are considerably higher than those of other baselines. The relatively high complexity of DRSNet mainly comes from RICA modules that combine the strength of Inception-ResNet and SENet. Although DRSNet has a high Flops, the Params and Param_size of DRSNet are still considerably less than those of other baseline CNNs. For example, AlexNet has a low computational complexity $(9.1 \times 10^7)$, but its Params and Param_size are approximately five times than those of DRSNet. This finding strongly indicates that DRSNet is a low-redundancy architecture regarding network parameters.

5. Conclusion

This paper introduces a novel CNN architecture called DRSNet for RS image recognition. Given that high-resolution commercial RS images are not practical for several communities, a Landsat 8 data set that could be used for various real-world applications is proposed. Our proposed network exhibits excellent classification results on Landsat 8 and three public RS data sets, namely, EuroSAT, Brazilian Coffee Scenes, and UCMerced Land Use. DRSNet outperforms baseline CNNs on our Landsat 8 by approximately 2%–9% in terms of OA and Kappa. It also achieves the highest classification accuracies on Brazilian Coffee Scenes and UCMerced. Although the OA and Kappa of DRSNet are slightly outperformed by Inception-ResNet-v2 on EuroSAT, it achieves the highest classification accuracies of four subclasses. The satisfactory performance of DRSNet indicates its strong classification ability and generalization capacity. Our effort could be pivotal to research and applications, such as land use and cover classification, urban planning, and environmental simulations.

Future studies will focus on investigating the principles and rules for building an effective CNN structure. Research from the computer science domain has revealed that several key factors, including the depth, width, and cardinality of networks, are crucial for CNN performance (Tan and Le, 2019; Xie et al., 2017). However, selecting and utilizing these factors...
properly when designing a CNN architecture remain an immense challenge for researchers outside the computer science domain. The findings of this paper will be used to conduct further experiments to investigate the influence of these factors on model performance. Rules of thumb will be summarized for RS researchers, which could be beneficial in narrowing down this gap.

CRediT authorship contribution statement

Fehai Chen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. Jin Yeu Tsou: Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


