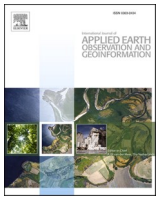
International Journal of Applied Earth Observations and Geoinformation 103 (2021) 102515

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A hybrid deep convolutional neural network for accurate land cover classification

Naftaly Wambugu [a], Yiping Chen [a,*], Zhenlong Xiao [a], Mingqiang Wei [b], Saifullahi Aminu Bello [a], José Marcato Junior [c], Jonathan Li [a,d,*]

[a] Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, FJ 361005, China
[b] School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[c] Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul, Campo Grande 79070900, Brazil
[d] Department of Geography & Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

ARTICLE INFO

Keywords:
Deep learning
very high spatial resolution (VHSR)
Remote sensing
Dilated convolution
Deep supervision
Spatial attention

ABSTRACT

Land cover classification provides updated information regarding the Earth's resources, which is vital for agricultural investigation, urban management, and disaster monitoring. Current advances in sensor technology on satellite and aerial remote sensing (RS) devices have improved the spatial-spectral, radiometric, and temporal resolutions of images over time. These improvements offer invaluable chances of understanding land cover information. However, land cover classification from RS images is an intricate task because of the high intra-class disparities, low inter-class similarities, and image variation types. We propose a cascaded residual dilated network (CRD-Net) for land cover classification using very high spatial resolution (VHSR) images to address these challenges. The proposed hybrid network follows the encoder-decoder concept with a spatial attention block to guide the network on learnable discriminate features coupled with an intermediary loss to enhance the training process. Moreover, a cascaded residual dilated module increases the network's receptive field to enrich multi-contextual features further, thus boosting the resultant feature descriptor. Extensive experimental results demonstrate that the proposed CRD-Net outperformed state-of-the-art methods, achieving an overall accuracy (OA) of 90.73% and 90.51% on the ISPRS Potsdam land cover dataset and ISPRS Vaihingen dataset, respectively.

1. Introduction

Earth observation imagery plays an essential role in developing accurate and timely thematic maps for land cover. It provides a precise understanding of anthropogenic processes on Earth's surface that is consistent and spatially continuous for a different range of spatial resolutions and time scales (Xiang et al., 2019). Thematic maps are mainly derived from the classification of RS images, which is effectively achieved through computer-aided analysis. Advances in sensor technology on satellites and aerial RS over the years have caused increasingly massive, accessible, and affordable imagery with high spatial and temporal resolutions. Besides, improved image quality and quantity, coupled with massive, accessible, and affordable computation power through the graphics processing units (GPUs) and parallel computing platforms, have led to superior computer algorithms. This, in turn, has inspired improvement in image analysis tasks such as scene understanding, detecting and segmenting objects, and pixel-level image classification (Xu et al., 2019).

Pixel-level image classification is a vital process in land cover mapping that assigns every image's pixel to a predefined class label where same-labeled pixels possess similar characteristics. VHSR image classification has various applications, such as mapping land use and land cover (LULC) (Weigand et al., 2020), vegetation classification (Flood et al., 2019), tracking watercourses and water bodies (Mishra et al., 2020; Pereira et al., 2019), urban ecology monitoring and understanding (Alshehhi and Marpu, 2021; Li et al., 2018), among others.

Correct and updated information about the land cover is essential for classifying, planning, predicting, tracking, and formulating ways to use the Earth's resources better and for the greater interest of humanity (Huang et al., 2019; Ojha et al., 2019; Yin et al., 2014). Solving land cover classification can help to overcome many obstacles relating to urban planning, environmental engineering, and natural landscape

* Corresponding authors at: Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, FJ 361005, China (Y. Chen and J. Li).
E-mail addresses: chenyiping@xmu.edu.cn (Y. Chen), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.jag.2021.102515
Received 24 July 2021; Received in revised form 18 August 2021; Accepted 20 August 2021

monitoring, among other applications (Huang et al., 2018).

Despite the great opportunities that land cover classification offers, classifying VHSR imagery poses a significant challenge due to the images' heterogeneity property, multi-object class imbalance, and varied distribution. Besides, effective classification of VHSR demands massive computation power, superior and robust algorithms with greater accuracy. Traditional methods of collecting and classifying land cover data are human-dependent, less efficient, and demanding on time and cost (Sang et al., 2020).

In this work, a novel cascaded residual dilated network (CRD-Net) is proposed to handle the intricate task of land cover classification using VHSR images. Our proposed framework is validated on the ISPRS Potsdam and Vaihingen land cover datasets by attaining competitive classification results in the two datasets without any post-processing strategy. The following is the outline of our work's contribution:

1) We propose a hybrid network for land cover classification using VHSR images, which uses spatial attention blocks to improve feature learning ability by focusing on essential learnable features; coupled with an intermediary loss function that enhances the network training procedure.
2) The proposed cascaded residual module enlarges the receptive field thus improving the network's multi-scale inference and elevates contextual information without falling into a gridding problem.
3) Extensive experiments on ISPRS 2D Postdam semantic labeling dataset and the Vaihingen dataset show that our proposed CRD-Net outperforms previous methods for land-cover classification.

Our work is presented as follows: Section 2 highlights the related work, indicating our novel contribution. Section 3 presents the study areas and dataset descriptions, the methods are described in Section 4. Experiments and discussions are presented in Section 5. Finally, conclusions are highlighted in Section 6.

2. Related work

Recent image processing tasks have witnessed great advances. This advancement comes from the increased image datasets and more available computation power through high-powered GPUs that have continuously facilitated the training of superior DCNNs for image analysis tasks (Cheng et al., 2017; Guo et al., 2017).

Image classification often relies on a larger receptive field's (RF) ability to obtain high-level features. Larger receptive fields ought to capture long-range semantic features to classify large objects correctly. Additionally, the low-level feature's fine spatial details are critical for optimal pixel-level classification. Several works have attempted exploiting semantic features extracted by DCNNs extensively. FCN (Sherrah, 2016) proposed a fully convolutional network without pooling layers for dense labeling of high-resolution aerial images, thus avoiding deconvolution or interpolation operations to overcome spatial information loss. SegNet (Badrinarayanan et al., 2017) utilized the encoder-decoder network for pixel-wise classification using pooling indices on the encoder phase and up-sampling on the decoder, while PSPNet (Zhao et al., 2017) employs feature pyramid pooling to achieve context aggregation by enlarging the kernel size.

In contrast, DeepLab+ (Chen et al., 2018), also called atrous spatial pyramid pooling (ASPP), exploited parallel dilated convolutions using varying dilation rates to probe multiscale image features. DeepLabv3+ (Chen et al., 2018) extends previous DeepLab versions by proposing the depthwise separable convolution in the ASPP and decoder modules, resulting in a more efficient and robust encoder-decoder based network. Whereas these methods yield good results in various tasks, they may not adaptively capture all valid features necessary for pixel-level classification for VHSR images (Liu et al., 2019).

In the land cover classification tasks, RefineNet (Lin et al., 2017) used a multi-path refinement approach to explicitly exploit the information from the downsampling phase to attain original image resolution using long-range skip connections. This overcame repeated downsampling operations such as pooling or convoluted strides, which lead to a reduction in the initial image's spatial resolution. GFRNet (Islam et al., 2017)proposed memory gates between the layers to handle multiscale contexts to optimize the selection criteria of pixels forwarded in the DCNN network,while SCAttNet (Li et al., 2021), employed spatial and attention mechanisms to refine features adaptively using a lightweight network. Whereas using gates can filter the unnecessary features from passing through the network, it can slow down the network and may not handle complex VHSR data satisfactorily. In other works, generative adversarial networks (GANs) and conditional random fields (CRF) were combined to refine classification maps for hyperspectral image classification (Zhong et al., 2020).

Dilated convolution (DC) (Yu and Koltun, 2016), has become a core approach in many multi-class segmentation tasks due to its power in multi-contextual feature aggregation without loss of spatial information. As a result, DC has been explored in various image analysis and classification tasks (Duarte et al., 2018; Hamaguchi et al., 2018; Zhou et al., 2018).

Also, notable success in the application of attention mechanisms in natural language processing, has greatly inspired its broader adoption for image analysis tasks. For example, Wang et al. (2017) obtained more image discriminative feature representation by stacking attention modules to form attention-aware features for image classification, while Zhao et al. (2018) employed a bi-directional information propagation path to aggregate long-range contextual information using a point-wise spatial attention mechanism, that helped in fusing global and local information to understand complex natural scenes better.

Following the intuition of (Liu et al., 2021), we propose a hybrid architecture that progressively learns more discriminative features while integrating complementary features in each network stage. Moreover, inspired by (Lee et al., 2015), we employ intermediary loss at the intermediate layers of the encoder sub-network to improve the training process and promote deeper supervision. The proposed architecture exploits the power of the attention mechanism in precise feature learning and exploits extensive information flow between the nested dilated layers to fully exploit multi-scale contexts.

3. Study areas

In this work, ISPRS 2D Semantic Labeling Contest - Potsdam and ISPRS Vaihingen datasets from the International Society for Photogrammetry and Remote Sensing (ISPRS) (Rottensteiner et al., 2012) are used. The standard benchmark datasets comprise aerial images over the urban area of Potsdam city and the Vaihingen region in Germany. The Vaihingen region dataset is obtained from a city with many separate buildings and smaller villages with several separate multi-story buildings. Meanwhile, the Potsdam city dataset comprises a historical city whose building blocks are vast and dense with narrow streets. Each dataset contains six labeled land cover classes that are most popular.

The six categories have been defined as impervious surfaces, buildings, low vegetation, trees, cars, and clutter with the white, blue, cyan, green, yellow, and red color codes respectively. The clutter class includes some water bodies and other incoherent objects from the task. Fig. 1 shows the localization of study areas of the two datasets.

3.1. Potsdam dataset

The Potsdam land cover dataset was developed to enhance automated delineation of urban objects from RS data. This dataset contains very high-resolution objects which are heterogeneous, thus making the classification task quite challenging. The dataset is focused on elaborate 2D per-pixel labeling on multiple classes. It seeks to support scientific methods and superior models working towards full automation for 2D object recognition and image classification. The Potsdam dataset

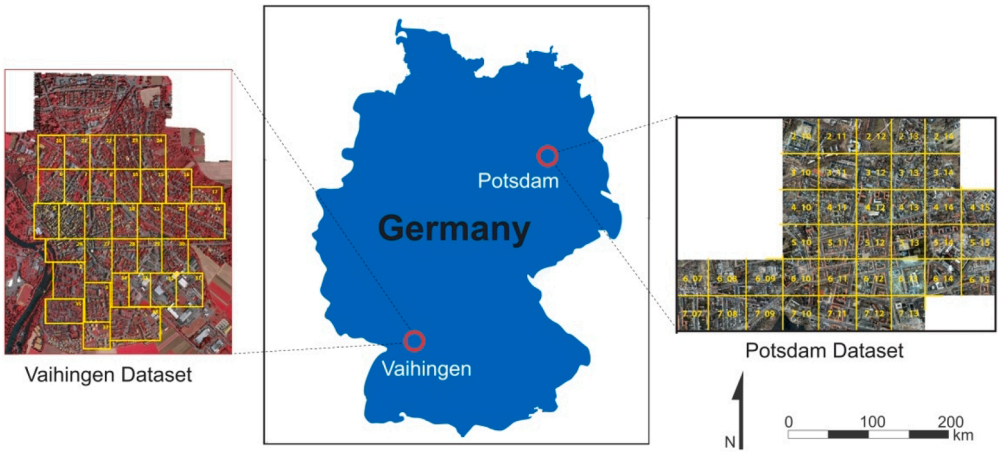


Fig. 1. Study areas of the Potsdam dataset and the Vaihingen dataset (Rottensteiner et al., 2012).

contains 38 tiles of 6,000 px × 6,000 px with 5 cm ground resolution. We used 14 of such images as test images and used only RGB images in our experiments.

3.2. Vaihingen dataset

The ISPRS 2D Semantic Labeling Challenge provides the Vaihingen dataset for image classification and 2D labeling. The dataset contains 33 image tiles of 2,494 px × 2,064 px with 9 cm ground resolution. 16 of the 33 tiles have been labeled. Only near-infrared, red, and green (IRRG) bands were used in our experiments.

4. Proposed method

In this work, a hybrid network called CRD-Net is proposed, which comprises the following components: a) an encoder-decoder sub-network to recover the lost spatial details caused by down-sampling operations with dual spatial attention blocks to guide the network in focusing on essential features; b) intermediary loss function connected to the spatial attention blocks for improved feature learning. c)the CRD module to attain better multi-scale contextual response and information flow between the layers. Each component of the proposed network is discussed in the later sections and the pipeline of the CRD-Net architecture is presented in Fig. 2.

4.1. Encoder-decoder with a dual spatial attention mechanism

The encoder-decoder paradigm is common for image classification networks owing to its ability to probe image features and harness the required high-level discriminative information (Zhou et al., 2018). In

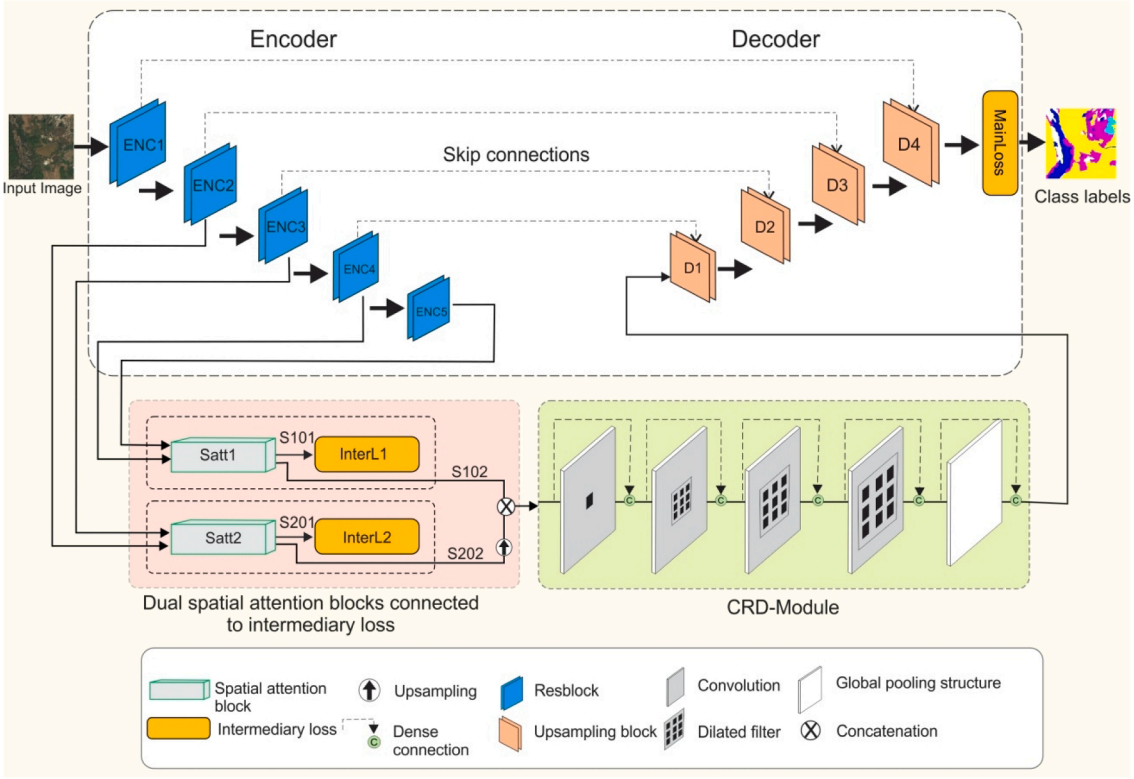


Fig. 2. The proposed Cascaded Residual Dilated Network Architecture.

addition, the decoder subnet's ability to recover spatial details lost through pooling or stride convolution operations on the encoder subnet makes it preferred for most semantic segmentation tasks (Chen et al., 2018). In general, the encoder-decoder network contains an encoder subnet that successively shrinks the feature maps and exploits high-semantic details as the network gets deeper. On the other hand, the decoder subnet recovers the crucial spatial details lost in the encoder segment of the network.

Based on this idea, our network architecture uses an encoder with a pretrained ResNet-101 network with five residual blocks marked ENC1 to ENC5 as the backbone, which works as an effective feature extractor. We use a pretrained network, since utilizing pretrained parameters and weights can help reduce the massive training data required to train deep networks from scratch and facilitate faster model convergence (Briechle et al., 2021).

We employ two spatial attention blocks named Satt1 and Satt2 in our network architecture to generate spatial feature maps from spatial relationships of the features as shown in Fig. 2. Fig. 3 shows The spatial attention block aims at learning a weight map representing the relative importance of activation for the spatial dimensions. The convolutional image feature maps from ENC2 and ENC3 are branched out into three copies representing the key, value, and query as illustrated in Fig. 3. The three copies are represented as f(x), h(x), and g(x), respectively to form the attention block named Satt1. We then apply the dot product attention to generate the resultant attention feature maps. Equally, the convolutional feature maps for ENC4 and ENC5 are branched out into three copies of f(x), h(x), and g(x), representing the key, value, and query values in the second spatial attention block named Satt2. The two spatial attention blocks are connected to two intermediary losses labeled InterL1 and InterL2 through outputs S1O1 and S2O1. The up-sampled output of Satt1 is concatenated with the output of Satt2 and then connected to the CRD module in a cascaded fashion. This helps the model focus selectively on discriminative features and ignore redundant and less important information (Vaswani et al., 2017). Besides, weighting the channels of the feature maps selectively can significantly improve the feature learning in residual modules and have shown significant improvement in semantic segmentation tasks (Zhong et al., 2020). Our networks extensively exploit residual connections in the encoder-decoder subnetwork and the CRD module owing to the success of deep residual networks in image classification for both spectral and spatial data (Zhong et al., 2018).

4.2. Deep supervision with intermediary loss

Loss functions inform how erroneous the classification prediction is from the ground truth. This is achieved through backpropagation. Most deep ConvNets implement loss function at the output layer, where the loss is propagated backward to earlier layers. However, single supervision at the output layer may not adequately learn complex features in the hidden layers, resulting in classification errors (Liu et al., 2020).

To better evaluate the loss in earlier layers and supervise the network, a loss objective is introduced at the intermediate layers of the deep neural network to improve the learning process of hidden layers making it more transparent and direct (Lee et al., 2015). This follows the intuition that a discriminative classifier working as a proxy can learn high discriminative features from hidden layers of the network and can better provide inference during hidden middle layers weight updates. Besides, the intermediary loss function introduced at intermediary layers of the network can significantly improve the supervision in deep networks at the layer level compared to relying on the results of the backpropagation process from the output layer (Muhammad et al., 2018).

Following (Zhao et al., 2017), two intermediary losses, InterL1 and InterL2, are introduced at the output of spatial attention Satt1 and Satt2 to effect direct supervision in the intermediate layers, as shown in Fig. 2. The learning process is decomposed where the two intermediary losses pass through the intermediate layers, thus optimizing the network training process.

We derive a multi-task loss function (L_{Total}) by combining the weights of the three losses as defined in Equation (1).

$$L_{Total} = \sum_{i=1}^{t} \lambda_i L_i \tag{1}$$

where L_i is the loss and λ_i is the weight associated with task i.

Backpropagation seeks to achieve convergence at the least loss weight value; different loss weights for different tasks can be distributed across several tasks, with each task having a significant influence on the network training (Chennupati et al., 2019) . Besides, adaptive tuning of task weights can help optimize the learning process. To achieve this, we employ dynamic weight adjustments (Guo et al., 2018) to update the weights α, β, and γ. The total loss in our network is defined by

$$Total\ Loss = (\alpha \times MainLoss) + (\beta \times InterL_1) + (\gamma \times InterL_2) \tag{2}$$

where α, β, and γ are the respective weights in our network, and InterL1, InterL2, and MainLoss are loss values of the output layer, the Satt1 spatial block, and Satt2 spatial block, respectively. By feeding the output of the attention mechanism to the intermediate loss functions, the proposed network can learn features more precisely guided by the intermediary loss that enhances deeper supervision at the intermediate layers of the network.

We scale the weights α, β, and γ for intermediary losses InterL1,

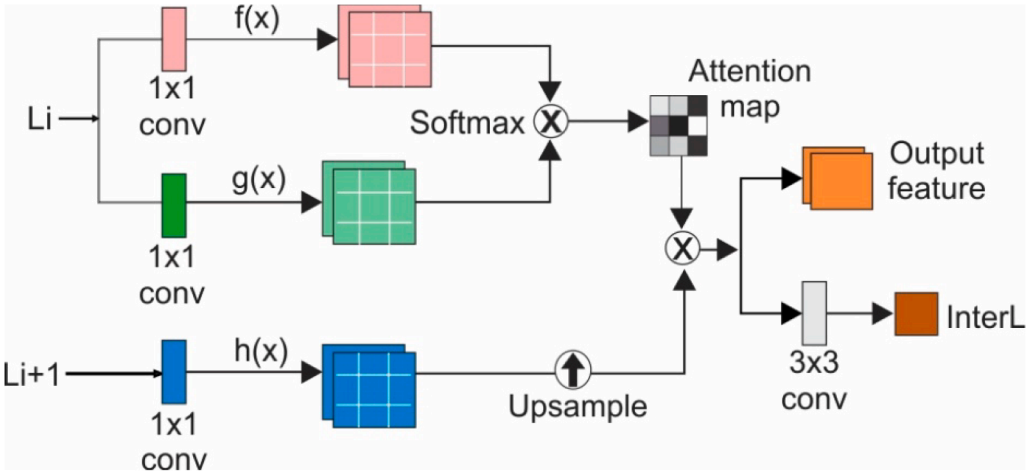


Fig. 3. Spatial-attention block used in the CRD-Net.

InterL2, and the MainLoss respectively to evaluate their influence on the network's training process, as highlighted under the ablation studies section. The three losses InterL1, InterL2, and the MainLoss cumulatively, contribute to the final prediction.

Weighted cross-entropy (WCE) loss (Martinez and Stiefelhagen, 2018) sums up pixels loss in a given mini-batch. In datasets with a high variation of pixels per class on the training set, class balancing is applied where the loss is weighted differently based on the true class. In this case, weights for classes with fewer pixels are elevated, while weights for classes with more pixels are diminished. We use weighted cross-entropy loss with median frequency balancing (MFB) (Eigen and Fergus, 2015). The WCE loss is defined by

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} W_i \; p_i \; log\left(\frac{e_{pi}}{\sum_{j=1}^{N} e^{pj}}\right) \tag{3}$$

where N represents the commulative classes, W_i denotes the class weight i, p_i, represents the prediction while $\acute{p}_i$ is the ground truth distribution of class i.

4.3. Cascaded residual dilated module

We propose the CRD module to progressively enlarge the RF in a scalable manner to help our network capture multi-scale contextual information and enhance elaborate information flow without increasing the network complexity or falling into gridding problem.

Dilated convolution helps aggregate multiscale contextual details without sacrificing the image resolution. The CRD module consists of dilated convolutional layers with no pooling or subsampling and gradually expands the RF without resulting in loss of spatial resolution or coverage. The dilated kernel with parameter $r > 1$ causes an enlargement in the RF without raising the number of parameters or computation requirements; different rates can be set to adjust the receptive field range. A standard dilated convolution is obtained as defined in Equation (4)

$$F = (r-1)(k-1)+k \tag{4}$$

where r denoted the dilation rate, and k represents the kernel size, and F represents the receptive field.

For a standard convolution operation with $k \times k$ kernel, S denotes the stride, which can have the following instances: $S > 1$, implies a down-sampling operation, $S = 1$, maintains the resolution of the feature map (considering adequate padding), and $0 < S < 1$, implying up-sampling which increases the feature map size. Enlarging a kernel of $k \times k$ filter to a kernel of $k + (k\text{-}1)\ (r\text{-}1)$, with r representing the dilated stride allows flexible aggregation of multiscale contextual details from the receptive field while maintaining the same resolution dimensions.

Although the CRD module presents great benefit in expanding the receptive field, it can generate holes called gridding artifacts (Chen et al., 2018; Yu and Koltun, 2016), where neighboring output units are processed from separate input sets resulting in different actual RF. This implies that some kernel responses do not act on some regions in the receptive field causing variability in kernel responses from the receptive field. To cure the gridding problem, the proposed CRD module progressively concatenates the residual connections with the resultant and previous feature maps of cascaded dilated layers. The resultant improved information flow between the dilated convolutional layers ensures that all kernel responses are obtained from the full receptive field thus overcoming the gridding problem.

Moreover, since image classification of VHSR images requires a descriptor with sufficient short, medium, and long-range semantics, the residual connections enhance information flow and boost significant features between the dilated layers following the intuition of (Wang et al., 2019). The proposed CRD module is illustrated in Fig. 4. Each layer receives the feature map from the two concatenated spatial attention blocks as input and performs a dilated convolution operation with rates of r_1, r_2, r_3, and r_4. Through residual connections, the resultant and previous feature maps of every dilated layer are combined. By gradually increasing the dilation rate in the cascading layers, the network is robust to achieve an effective full receptive field where lower dilation rates obtain fine details and small objects spatial dimensions, while larger dilation rates capture the larger objects' features resulting in a robust feature descriptor.

The hierarchical fusion of all the layers from smaller to larger dilation rates allows the participation of the dilated convolution layer pixels in probing the multiscale features before concatenation. This ensures information sharing between the layers, thus overcoming the gridding problem.

The CRD model is designed to enlarge the receptive field with fewer parameters based on the dilation convolution. The final feature map is generated from the computation of all features after aggregating the receptive fields of each layer and effectively capturing the multi-scale contextual information. All the feature vectors from different dilated layers are concatenated in the global pooling layer before inputting to the decoder unit.

4.4. Quantitative assessment measures

We use precision (PRE), recall (REC), intersection over union (IoU), pixel accuracy (PA), and mean F1-score (mF1) to evaluate the performance of our proposed method.

$$PRE = TP\%(TP + FP) \tag{5}$$

$$REC = T\ P\%(TP + FN) \tag{6}$$

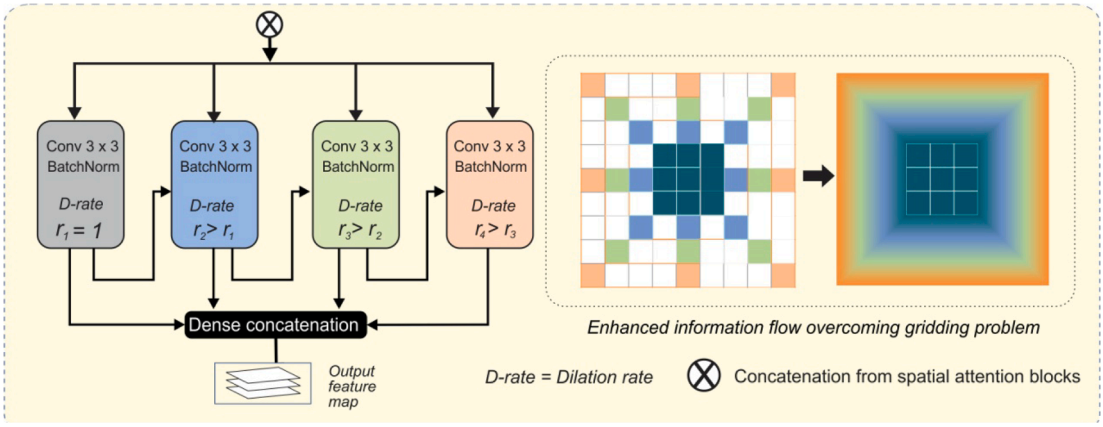


Fig. 4. Illustration of 3×3 CRD module with dilation rates of r_1, r_2, r_3, and r_4 and the resultant nested kernel.

$$IoU = TP\%(TP + FN + FP) \quad (7)$$

$$PA = (TP + TN)\%(TP + TN + FP + FN) \quad (8)$$

$$F1 - Score = 2 \times (Precision \times Recall)\%(Precision + Recall) \quad (9)$$

where TP, TN, FP, and FN, denotes true positive pixels, true negative pixels, false positive pixels, and false negative pixels, respectively.

5. Experiments and discussions

In this section, the network's training details are explained first. Second, the data pre-processing procedures and the performance evaluation are presented. Third, the ablations experiments results are discussed.

5.1. Training details

We implemented our method using Pytorch deep learning framework. The basic experiment platform is on NVIDIA GeForce RTX 2080Ti GPU graphics card, which contains 11 GB memory and is equipped with CUDA 10.2 and Cudnn7.6.5. Both the training and testing are carried out on this platform.

We adopt the settings used in the related work on dilated convolution (Liu et al., 2019). We trained our model using Adam optimizer (Kingma and Ba, 2014) with AMSGrad (Reddi et al., 2018) and weight decay of 2×10^{-5}. The polynomial learning rate (LR) decay is set as $(1 - (\text{cur iter}/\text{max iter}))^{0.9}$. The input image size is 256×256 pixels with a batch size of 5. Images 24, 2, 3, and 14 are set for the training, validation, local test set, and hold-out test set for Potsdam dataset images while images 16, 2, 4, and 17 are similarly set for the Vaihingen dataset.

5.2. Pre-processing and augmentation

In the land cover segmentation task, image pre-processing can speed up the network's self-fitting process more effectively (Li et al., 2020). Since the training samples are not sufficient enough, we adopt data augmentation (Wong et al., 2016) as an effective method of supplementing the training samples. To achieve this, we employed the albumentations library (Buslaev et al., 2020), which provides several flexible and efficient image augmentation functions relating to color, contrast, brightness, and other geometric transformations.

5.3. Performance evaluation

We demonstrated the efficacy of the CRD-Net architecture by carrying out experimentson the two datasets. The per-class classification results on the test sets of the Potsdam (RGB) dataset and Vaihingen (IRRG) datasets trained separately using the CRD-Net model are shown in Tables 1 and 2 The average scores are computed from all classes, but the clutter.

The clutter class is ignored due to its minor representation in the sample distribution of the training set. Our network achieved an overall mF1-score of 92.1% and 90.0% on the Potsdam and Vaihingen datasets, respectively. The visual classification results of the proposed model on Potsdam and Vaihingen datasets are presented in Figs. 5 and 6, respectively.

Table 1
Per class results (in percentage) on test set of ISPRS Potsdam dataset trained with CRD-Net model.

Metrics	Buildings	Trees	Low Veg.	Clutter	Road Surface	Cars	Avg
Precision	88.9	87.6	88.8	61.7	94.4	89.0	89.8
Recall	97.6	89.7	84.2	62.0	88.4	98.0	91.6
$_{F1}$-*Score*	96.7	88.6	87.4	63.9	92.9	94.8	92.1
IoU	87.0	79.6	76.1	44.8	84.0	87.4	82.8

Table 2
Per class results (in percentage) on the test set of ISPRS Vaihingen dataset trained with CRD-Net model.

Metrics	Buildings	Trees	Low Veg.	Clutter	Road Surface	Cars	Avg
Precision	95.0	86.8	85.2	89.4	93.4	88.1	89.7
Recall	95.7	92.6	81.7	42.7	92.0	89.2	90.3
F1-Score	95.4	89.6	83.5	55.8	92.7	88.7	90.0
IoU	91.1	81.1	71.6	40.6	86.4	79.6	82.0

Tables 3 and 4 show our network's OA and mF1-score performance compared with other networks on the land cover classification task on the two benchmark datasets. Our network achieves an OA score of 90.7% and an mF1-score of 92.1% on the Potsdam dataset outperforming the other compared models. Comparing our proposed network's performance on the Potsdam dataset using OA, the CRD-Net achieves achieved 0.4% higher compared to the FCN based DST 2 (Sherrah, 2016), which is the second-best. Compared with SCAttNet (Li et al., 2021), based on spatial and channel attention, our model achieves a 5.2% score higher. This implies that both deep supervisions through intermediate loss function and the CRD model improve the overall classification accuracy. Moreover, comparing CRD-Net with DeepLab3+ (Chen et al., 2018), based on atrous pyramid pooling, our network achieved a superior performance of 5.5% higher. In addition, CRD-Net improves SegNet (Badrinarayanan et al., 2017) and RefineNet (Lin et al., 2017) by 7.9% and 7.3%, respectively.

The CRD-Net achieved an OA score of 90.51% and an mF1-score of 90.0%, on Vaihingen dataset, outperforming all other models compared in the study. Moreover, the CRD-Net model attained the best mF1 per class scores on all categories except for road surface class. Compared with the dense dilated convolution merging DDCM network, our model achieved a marginal improvement of 0.1%, with marginal improvement in all classes except for the road surface. The CRD-Net achieved an improvement of 3.7% OA on gated refinement network G-FRNet V2. Moreover, compared with the other models, our network improves SegNet (Badrinarayanan et al., 2017), DeepLabV3+, and RefineNet (Lin et al., 2017), by 10.2%, 3.7%, and 6.1%, respectively.

Notably, the growing demand for superior consistent computer-based analysis methods driven by the increased availability of very fine-resolution RS images calls for combined efforts from the research community to handle the cited challenges in processing VHR data for land cover classification. Our proposed hybrid network seeks to complement the existing methods proposed for handling the intricate task of land cover classification.

Specifically, our proposed method seeks to demonstrate the power of hybrid approaches in classification tasks by; blending the power of dilated convolution in harnessing contextual information for multi-sized objects with existing approaches like attention mechanism for and deep supervision that guides the training process using intermediate loss function.

The proposed method is new in land cover classification using VHR data. Although the method posted marginal improvement on the second-best method, it can motivate and spur more research into improved hybrid models for land cover classification problems with superior performance.

5.4. Ablation experiments

We analyzed the proposed network's performance by carrying out further experiments using various configurations to validate the significance of each sub-component of the proposed hybrid network. The per-class OA and mF1 scores of the CRD-Net under different configurations are presented in Table 5. Additionally, we present the visual

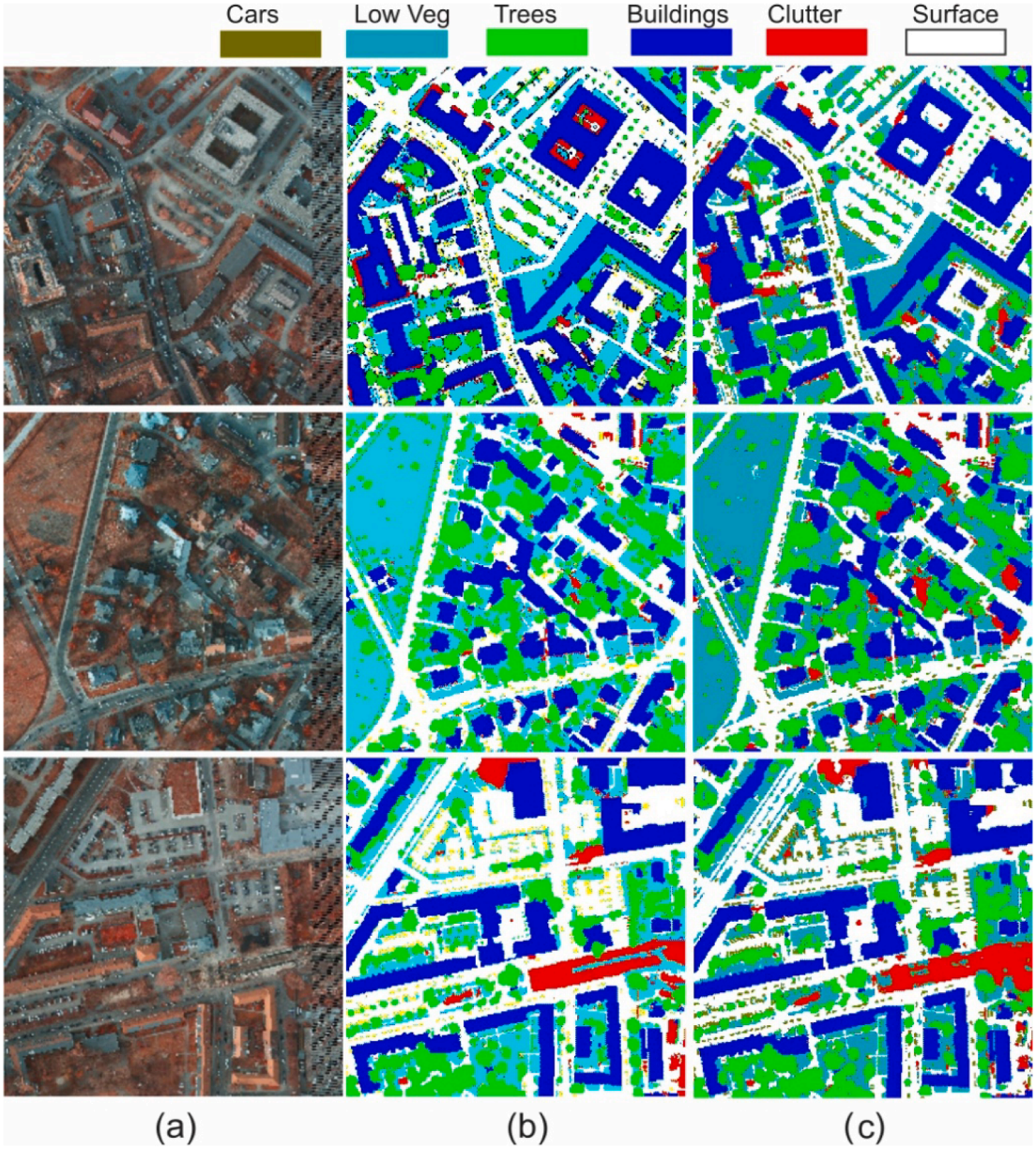


Fig. 5. Classification results for test images of Potsdam RGB data tiles 4_15, 2_12, and 6_15, respectively. (a) input image tile, (b) ground truth, and (c) prediction.

classification results of our proposed architecture under different network configurations in Fig. 7.

5.4.1. Evaluating the impact of the spatial attention module.

To better understand the impact of spatial attention mechanisms on the CRD-Net architecture, we carried out ablation experiments on the network without incorporating the spatial attention block in the network. The results show that the OA accuracy drops by 0.27% after dropping the spatial attention block from the network. Besides, the influence of spatial attention in focusing on specific regions in the layer is validated from the visualized classification results, where car class in the marked regions are viewed as detached in the CRD-Net model results. Still, the regions are visibly connected in the model with no spatial attention results. This demonstrates the ability of spatial attention in delineating regions, especially in high-resolution images in their natural setting where objects of dissimilar classes may possess similar features or in scenes where intra-class variation is present.

5.4.2. Evaluating the impact of deep supervision and intermediary loss

The effect of training with no intermediate loss is investigated. We notice that intermediary loss influences network performance. Specifically, training the network using weighted intermediate loss results in an improvement of 0.16% on the classification results. The intermediate loss functions introduced at middle layers help in guiding the training process resulting in improved accuracy. Using relative weights to compute the total loss improved the network performance since the main loss contributes more to the final prediction. When the network is trained using the same weights for all losses, the accuracy is dropped by 0.11%. We observe that integrating deep supervision at the intermediary layers can improve gradient flow, reduce the vanishing gradient problem and improve network convergence.

5.4.3. Evaluating the impact of the CRD module

Ablation experiments results without the CRD module show a drop in AO by 0.43% while incorporating the CRD module in the proposed network recorded improvement on the OA in all classes. In addition, the CRD module demonstrates improvement in the visual classification

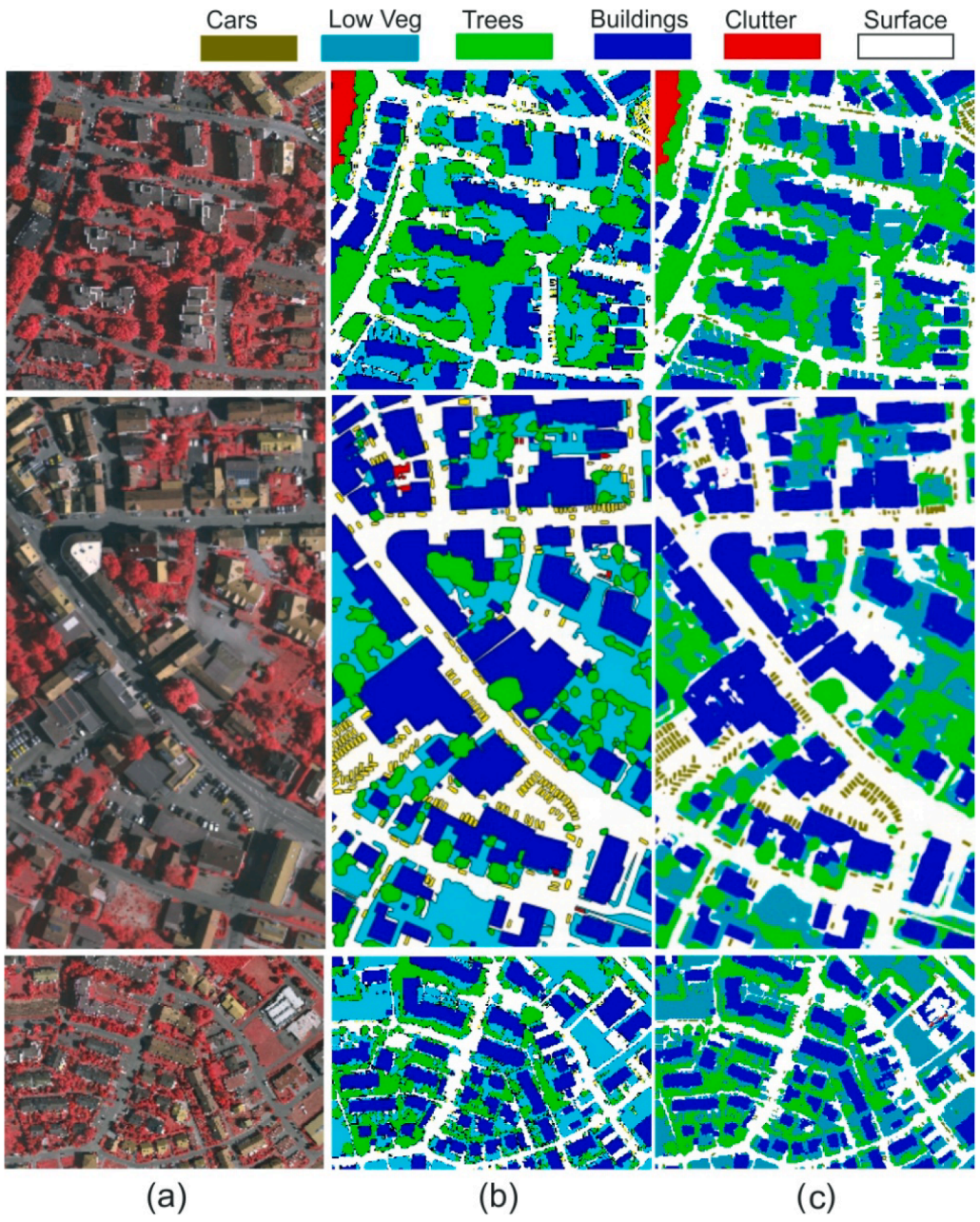


Fig. 6. Classification results for test images of Vaihingen IRRG image tiles *2, 27*, and *38* respectively. (a)input image tile, (b)ground truth, and (c) prediction.

Table 3
Comparison of OA and per class mF1(in percentage %) between CRD-Net and other published methods on ISPRS Potsdam land cover dataset.

Models	OA	Road Surface	Buildings	LowVeg	Trees	Cars	mF1
SegNet (Badrinarayanan et al., 2017)	82.9	86.2	88.3	77.2	80.0	54.2	77.0
DeepLabV3+ (Chen et al., 2018)	85.2	88.5	90.0	80.2	80.1	68.9	81.5
RefineNet (Lin et al., 2017)	83.4	87.3	86.9	78.3	79.6	75.9	81.6
SCAttNet V2 (Li et al., 2021)	85.5	91.2	90.3	80.0	80.3	70.5	82.1
DST 2 (Sherrah, 2016)	90.3	92.5	96.4	86.7	88.0	94.7	91.8
Ours	**90.7**	**92.9**	**96.7**	**87.4**	**88.6**	**94.8**	**92.1**

results in extracting pixel-level information effectively. By harnessing the capability of dilated convolution in enlarging RF to harness rich multi contextual information, the proposed network can delineate small objects precisely and capture large objects.

5.4.4. Combining the different components to form a hybrid network

Our proposed framework comprises of components collectively forming a hybrid network to handle the intricate task of land cover mapping. Since RS data contains many different sized objects whose features and settings are critical for land cover classification, classifying RS images becomes more challenging when some key objects are

Table 4
Comparison of OA and per class mF1 (in percentage %) between CRD-Net and other published methods on ISPRS Vaihingen land cover dataset.

Models	OA	Road Surface	Buildings	LowVeg	Trees	Cars	mF1
SegNet (Badrinarayanan et al., 2017)	80.3	81.1	86.4	78.0	73.9	85.7	81.0
DeepLabV3+ (Chen et al., 2018)	86.8	89.3	92.8	83.4	78.4	88.2	86.4
RefineNet (Lin et al., 2017)	84.4	87.6	88.5	81.9	79.1	87.9	85.0
G-FRNet V2 (Islam et al., 2017)	86.8	89.2	92.7	82.8	79.0	86.3	86.0
DDCM (Liu et al., 2019)	90.4	**92.7**	95.3	83.3	89.4	88.3	89.7
Ours	**90.5**	**92.7**	**95.4**	**83.4**	**89.6**	**88.7**	**90.0**

Table 5
Comparison of per class OA and mF1 (in percentage %) of CRD-Net with different configurations on ISPRS Vaihingen land-cover dataset.

Method	OA	Road Surface	Buildings	LowVeg	Trees	Cars	mF1
CRD-Net	90.5	92.7	95.4	83.4	89.6	88.7	90.0
Without Auxilliary Loss	90.4	92.8	95.0	83.3	89.7	88.2	89.8
Without Spatial attention	90.2	92.4	95.0	83.3	89.4	86.8	89.3
Without CRD module	90.1	92.3	94.9	82.7	89.3	88.3	89.7
CRD-Net With ResNet50	90.2	92.4	94.8	83.2	89.6	87.5	89.5

invisible, or suppressed due to their size, shadow, occlusion from the surrounding objects, or where the background suppresses the objects of interest. Besides, most RS data contains superfluous objects which can affect the accurate classification of land cover classes.

Since spatial information is indispensable for correct pixel-level classification, the proposed hybrid exploits the power of encoder-decoder to recover and restore spatial details suffered from down-sampling operations. Furthermore, the attention mechanism helps the network focus on key discriminative regions in the images by according such areas higher weights while suppressing redundant and less important regions such as backgrounds. Additionally, by utilizing intermediary loss functions, the model improves the learning process, where intermediate loss guides the backpropagation process, by defining how badly the network performs at the intermediary layers of the network as opposed to using a single loss function at the end of the network.

Finally, since rich and multi-scale contextual representation plays an essential role in correct classification of objects, especially with varied sizes such as in the case of VHRS images, we employ cascaded dilated convolutions to cause enlargement of the RF to obtain multi-contextual details without loss of spatial information.

The results in Table 5 shows that combining different components for the complex task of land classification can result in improved classification results. The hybrid network can learn more discriminative features progressively in each stage as complementary features are integrated, thus harnessing the benefits of each component.

However, models accuracy is greatly affected by the shadows cast by elevated objects such as buildings, trees, and vegetation thus causing great difficulty in VHR image classification tasks. The classification accuracy gets compromised if the objects' shadows are not detected and delineated during the classification process. Shadow detection, alignment, and correction for land cover classification using VHRS aerial imagery remains a great area of interest requiring further attention to mitigate shadow-prone errors.

Although the proposed hybrid network attained competitive classification results, obtaining optimum results by blending several components requires further attention on how best to integrate the sub-components for better prospects.

6. Conclusion

In this work, a hybrid network named CRD-Net is presented to tackle the challenging task of land cover classification with VHSR images. The proposed architecture harnesses short-range, mid-range, and long-range semantic information at different stages of the network while preserving the spatial details to generate a robust feature descriptor. The attention mechanism with intermediary loss at the encoder subnet assists in refined feature learning and attaining intermediary layers' deep supervision. Moreover, the network harnesses rich global multi-contextual information using the CRD module without falling into the gridding problem caused by dilation. Future experiments are necessary to validate the proposed framework for land cover mapping in other RS datasets and related tasks.

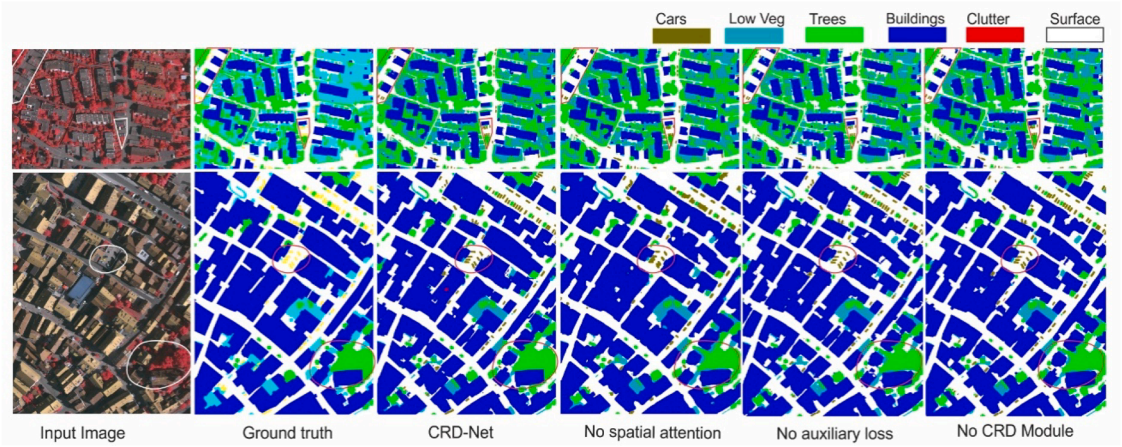


Fig. 7. Classification results for test images of Vaihingen IRRG image tiles 4 and 25, respectively, using different network configurations.

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.

Acknowledgement

The authors express profound gratitude to ISPRS for availing of the ISPRS benchmark datasets used in this research work.

Funding

This work was supported by the National Natural Science Foundation of China under grants of 41871380, and the Natural Sciences and Engineering Research to the Council of Canada under a grant of 50503–10284. The first author acknowledged the China Scholarship Council (CSC) to support his Ph.D. study at the School of Informatics, Xiamen University. Finally, the authors acknowledge CAPES PrInt (grant number 88881.311850/2018–01).

References

Alshehhi, R., Marpu, P.R., 2021. Extraction of urban multi-class from high-resolution images using pyramid generative adversarial networks. Int. J. Appl. Earth Obs. Geoinf. 102, 102379 https://doi.org/10.1016/j.jag.2021.102379.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Briechle, S., Krzystek, P., Vosselman, G., 2021. Silvi-Net – A dual-CNN approach for combined classification of tree species and standing dead trees from remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 98, 102292 https://doi.org/10.1016/j.jag.2020.102292.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and Flexible Image Augmentations. Inf. 11, 125.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proc. of ECCV. https://arxiv.org/abs/1802.02611.

Cheng, G., Han, J., Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. In Proc. of IEEE 105 (10), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.

Chennupati, S., Sistu, G., Yogamani, S., Rawashdeh, S., 2019. AuxNet: Auxiliary Tasks Enhanced Semantic Segmentation for Automated Driving. International Conference on Computer Vision Theory and Applications 645–652. https://doi.org/10.5220/0007684106450652.

Duarte, D., Nex, F., Kerle, N., Vosselman, G., 2018. Satellite Image Classification Of Building Damages Using Airborne And Satellite Image Samples In A Deep Learning Approach. In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2. ISPRS, pp. 89–96. https://doi.org/10.5194/isprs-annals-IV-2-89-2018.

Eigen, D., Fergus, R.. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. https://doi.org/10.1109/ICCV.2015.304.

Flood, N., Watson, F., Collett, L., 2019. Using a U-net convolutional neural network to map woody vegetation extent from high resolution satellite imagery across Queensland, Australia. Int. J. Appl. Earth Obs. Geoinf. 82, 101897 https://doi.org/10.1016/j.jag.2019.101897.

Guo, M., Haque, A., Huang, A., Yeung, S., Li, F.F., 2018. Dynamic Task Prioritization for Multitask Learning. In Proc. of ECCV. Springer, pp. 282–299. https://doi.org/10.1007/978-3-030-01270-0_17.

Guo, Y., Liu, Y., Georgiou, T., Lew, M.S., 2017. A review of semantic segmentation using deep neural networks. Int. J. Multimedia Inf. Retrieval 7, 87–93. https://doi.org/10.1007/s13735-017-0141-z.

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., Hikosaka, S., 2018. Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery. In Proc. of WACV 1442–1450. https://doi.org/10.1109/WACV.2018.00162.

Huang, B., Zhao, B., Song, Y., 2018. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 214, 73–86. https://doi.org/10.1016/j.rse.2018.04.050.

Huang, Z., Dumitru, C.O., Pan, Z., Lei, B., Datcu, M., 2019. Can a Deep Network Understand the Land Cover Across Sensors?. In: In Proc. of IGARSS, pp. 9847–9850. https://doi.org/10.1109/IGARSS.2019.8899080.

Islam, M., Rochan, M., Bruce, N., Wang, Y., 2017. Gated Feedback Refinement Network for Dense Image Labeling. In Proc. of CVPR 4877–4885. https://doi.org/10.1109/CVPR.2017.518.

Li, B., Su, W., Wu, H., Li, R., Zhang, W., Qin, W., Zhang, S., Wei, J., 2020. Further Exploring Convolutional Neural Networks' Potential for Land-Use Scene Classification. IEEE Geosci. Remote Sens. Lett. 17, 1687–1691. https://doi.org/10.1109/LGRS.2019.2952660.

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. In Proc. of ICLR. https://arxiv.org/abs/1412.6980.

Lee, C.-Y., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z., 2015. Deeply-Supervised Nets. In: In Proc. of. PMLR, pp. 562–570. In: http://proceedings.mlr.press/v38/lee15a.pdf.

Li, H., Qiu, K., Chen, L., Mei, X., Hong, L., Tao, C., 2021. SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. IEEE Geosci. Remote Sens. Lett. 905–909. https://doi.org/10.1109/LGRS.2020.2988294.

Li, Z., Zhou, C., Yang, X., Chen, X., Meng, F., Lu, C., Pan, T., Qi, W., 2018. Urban landscape extraction and analysis in the mega-city of China's coastal regions using high-resolution satellite imagery: A case of Shanghai, China. Int. J. Appl. Earth Obs. Geoinf. 72, 140–150. https://doi.org/10.1016/j.jag.2018.03.002.

Lin, G., Milan, A., Shen, C., Reid, I., 2017. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. In Proc. of CVPR. 1925–1934. https://doi.org/10.1109/CVPR.2017.549.

Liu, Q., Fu, M., Jiang, H., Gong, X., 2020. Densely Dilated Spatial Pooling Convolutional Network using benign loss functions for imbalanced volumetric prostate segmentation DDSP network with benign loss for prostate segmentation. Current Bioinformat. 15 (7), 788–799. https://doi.org/10.2174/1574893615666200127124145.

Liu, Q., Kampffmeyer, M., Jenssen, R., Salberg, A.-B.. Dense Dilated Convolutions Merging Network for Semantic Mapping of Remote Sensing Images. https://doi.org/10.1109/JURSE.2019.8809046.

Liu, X., Chen, Y., Wei, M., Wang, C., Gonçalves, W., Junior, J., Li, J., 2021. Building Instance Extraction Method Based on Improved Hybrid Task Cascade. IEEE Geosci. Remote Sens. Lett. PP,, 1–5. https://doi.org/10.1109/LGRS.2021.3060960.

Martinez, M., Stiefelhagen, R., 2018. Taming the Cross Entropy Loss. In Proc. of GCPR. https://doi.org/10.1007/978-3-030-12939-2_43.

Mishra, V., Limaye, A.S., Muench, R.E., Cherrington, E.A., Markert, K.N., 2020. Evaluating the performance of high-resolution satellite imagery in detecting ephemeral water bodies over West Africa. Int. J. Appl. Earth Obs. Geoinf. 93, 102218 https://doi.org/10.1016/j.jag.2020.102218.

Muhammad, U., Wang, W., Hadid, A., 2018. Feature Fusion with Deep Supervision for Remote-Sensing Image Scene Classification. In: In Prof. of ICTAI, pp. 249–253. https://doi.org/10.1109/ICTAI.2018.00046.

Ojha, S.K., Challa, K., Vemuri, M.K., Yarlagadda, N.S.V., Kumar, B.L.N.P., 2019. Land Use Prediction on Satillite images using Deep Neural Nets. In: In Proc. of ICCS, pp. 999–1003. https://doi.org/10.1109/ICCS45141.2019.9065698.

Pereira, F.J.S., Costa, C.A.G., Foerster, S., Brosinsky, A., de Araújo, J.C., 2019. Estimation of suspended sediment concentration in an intermittent river using multi-temporal high-resolution satellite imagery. Int. J. Appl. Earth Obs. Geoinf. 79, 153–161. https://doi.org/10.1016/j.jag.2019.02.009.

Reddi, S.J., Kale, S., Kumar, S., 2018. On the Convergence of Adam and Beyond. In Proc. of Int. Conf. on Learning Representations, ICLR. https://openreview.net/forum?id=ryQu7f-RZ.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS Benchmark on Urban Object Classification and 3d Building Reconstruction. In: ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. I-3. ISPRS, pp. 293–298. https://doi.org/10.5194/isprsannals-I-3-293-2012.

Sang, Q., Zhuang, Y., Dong, S., Wang, G., Chen, H., 2020. FRF-Net: Land Cover Classification From Large-Scale VHR Optical Remote Sensing Images. IEEE Geosci. Remote Sens. Lett. 17, 1057–1061. https://doi.org/10.1109/LGRS.2019.2938555.

Sherrah, J., 2016. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. https://arxiv.org/abs/1606.02585.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is All you Need, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, pp. 5998-6008.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.. Residual Attention Network for Image Classification. https://doi.org/10.1109/CVPR.2017.683.

Wang, L., Wang, S., Chen, R., Qu, X., Chen, Y., Huang, S., Liu, C., 2019. Nested Dilation Networks for Brain Tumor Segmentation Based on Magnetic Resonance Imaging. Frontiers in Neuroscience 13. https://doi.org/10.3389/fnins.2019.00285.

Weigand, M., Staab, J., Wurm, M., Taubenböck, H., 2020. Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data. Int. J. Appl. Earth Obs. Geoinf. 88, 102065 https://doi.org/10.1016/j.jag.2020.102065.

Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D., 2016. Understanding Data Augmentation for Classification: When to Warp?. In: In Proc International Conference on Digital Image Computing: Techniques and Applications. DICTA, pp. 1–6. https://doi.org/10.1109/DICTA.2016.7797091.

Xiang, J., Li, S., Xiao, K., Jianping, C., Sofia, G., 2019. remote sensing Quantitative Analysis of Anthropogenic Morphologies Based on Multi-Temporal High-Resolution Topography. Remote. Sens. 11, 1–20. https://doi.org/10.3390/rs11121493.

Xu, Y., Du, B., Zhang, L., Cerra, D., Pato, M., Carmona, E., Prasad, S., Yokoya, N., Hansch, R., Le Saux, B., 2019. Advanced Multi-Sensor Optical Remote Sensing for Urban Land Use and Land Cover Classification: Outcome of the 2018 IEEE GRSS Data

Fusion Contest. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 12, 1709–1724. https://doi.org/10.1109/JSTARS.2019.2911113.

Yin, H., Pflugmacher, D., Kennedy, R.E., Sulla-Menashe, D., Hostert, P., 2014. Mapping Annual Land Use and Land Cover Changes Using MODIS Time Series. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 3421–3427. https://doi.org/10.1109/JSTARS.2014.2348411.

Yu, F., Koltun, V., 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In Proc. of In. Conf. on Learning Representations, ICLR.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In Proc. of CVPR 6230–6239. https://doi.org/10.1109/CVPR.2017.660.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J., 2018. PSANet: Point-wise Spatial Attention Network for Scene Parsing. In Proc. of ECCV 270–286. https://doi.org/10.1007/978-3-030-01240-3_17.

Zhong, Z., Lin, Z.Q., Bidart, R., Hu, X., Daya, I.B., Li, Z., Zheng, W.S., Li, J., Wong, A.. Squeeze-and-Attention Networks for Semantic Segmentation. https://doi.org/10.1109/CVPR42600.2020.01308.

Zhong, Zilong, Jonathan, Li, Luo, Zhiming, Chapman, Michael, et al., 2018. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. IEEE Trans. Geosci. Remote. Sens. 56 (2), 847–858. https://doi.org/10.1109/TGRS.2017.2755542.

Zhou, L., Zhang, C., Wu, M.. D-LinkNet: LinkNet with Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. https://doi.org/10.1109/CVPRW.2018.00034.