IMPROVING LAND COVER CLASSIFICATION WITH A SHIFT-INVARIANT CENTER-FOCUSING CONVOLUTIONAL NEURAL NETWORK

Cong Luo[1], *Yuansheng Hua*[1,2], *Lichao Mou*[1,2], *Xiao Xiang Zhu*[1,2]

[1]Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2]Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

Convolutional neural networks (CNNs) are widely employed in remote sensing community. The CNN-based, also known as patch-based land cover classification method has gained increasing attention. However, this method very often requires the aid of post-processing, otherwise it is difficult to obtain accurate boundaries separating different land cover classes. In this paper, we discuss the reason of this phenomenon and propose a shift-invariant center-focusing (SICF) network to deliver more accurate boundaries to improve the patch-based land cover classification. The principle of SICF is calculating the class score from a center-focusing area based on a shift-invariant feature extraction module to calibrate prediction. We employ three modern CNNs to build corresponding SICF networks, the evaluation results indicate that compared with the conventional CNNs, the improvements made by SICF for delivering accurate boundaries in land cover classification are significant.

Index Terms— convolutional neural network, shift-invariance, class activation maps, land cover classification

1. INTRODUCTION

Empowered by the significantly enhanced performance within past few years, convolutional neural networks (CNNs) are widely recognized as a vital tool for land cover classification in remote sensing (RS) community [1, 2]. In order to obtain a land cover classification map for a given RS image using the CNN-based, also known as patch-based method, firstly, CNN classifies a series of contextual patches of the same size extracted from the image in the sliding window approach individually, then each predicted label is assigned to the center pixel of the corresponding patch [3, 4, 5].

However, a major deficiency in this classification mechanism arises, it is difficult to obtain accurate boundaries separating different classes in the final classification map if the patch-based method is applied solely without the aid of post-processing (e.g., various segmentation technologies [3, 6]). To provide an insight into this problem, if an input patch contains contextual information correspond to more than one classes (e.g., both grassland and built-up area present in the same patch), which occurs frequently near the boundary areas, the prediction could be the class that does not appear in the center of the patch as illustrated in Fig. 1. Because CNN chooses the category associated with highest score after the logistic regression / softmax to be the predicted class, without specifically considering whether the corresponding features present in the center area of the input.

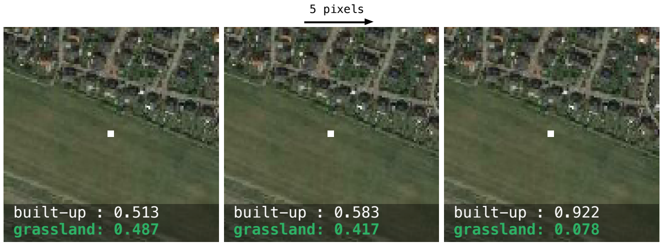

Fig. 1: Generating contextual patches for land cover classification by sliding window near a boundary area, each patch contains two land cover classes: grassland and built-up area. The top-2 predictions and corresponding probabilities output by ResNet-18 are listed accordingly. As the center points of the three contextual patches lay on the grassland area as represented by white boxes (enlarged for higher visibility), the correct result should be grassland (green label). However, all the three images are misclassified as built-up. Assigning the incorrect prediction to the center pixel leads to inaccurate boundaries.

In this work, we propose a shift-invariant center-focusing (SICF) network for providing a solution to deliver more accurate boundaries in the final classification map without further post-processing. With the help of class activation maps (CAM) which enabled the class-discriminative localization introduced in [7], the SICF network accommodates a calibration module that calibrates the initial prediction by calculating per-class score from a center-focusing area. Moreover, we noticed that there is increasing concern about the poor shift-invariance capability of modern CNNs, meaning a slight shift could lead to drastic differences in feature maps [8, 9]. Given that the calibration module relies on feature maps in the last convolutional layer to perform correctly, a shift-invariance

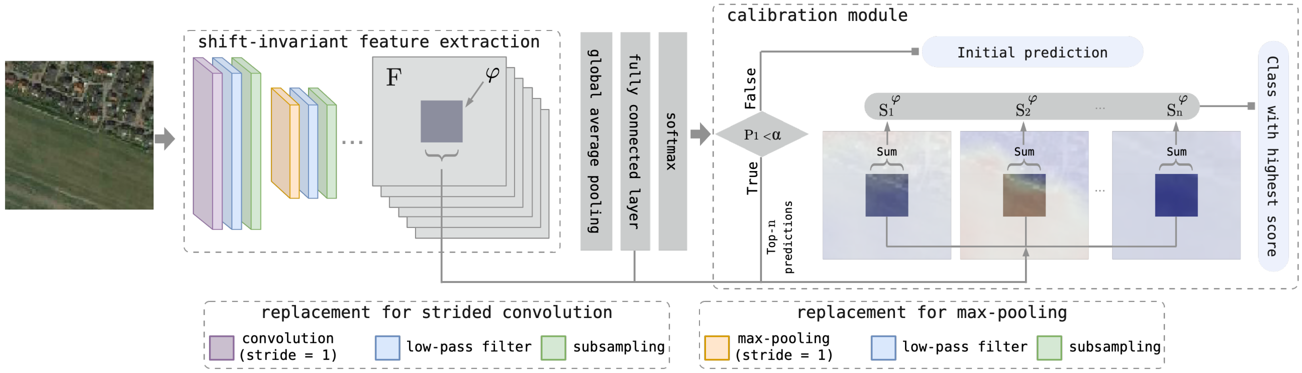

Fig. 2: Architecture of shift-invariant center-focusing (SICF) network.

improved feature extraction solution is integrated in the proposed network to provide a robust feature extraction during the window sliding. Finally, we apply the proposed method to three modern CNNs and test them on a dataset encloses six land cover classes to evaluate the performance.

2. METHODOLOGY

2.1. Calibration Module

In SICF network, a calibration module is affixed after the softmax layer. The main function of the calibration module is to let CNN eventually focus on the contextual information presents in the center area of the input image. The principle is inspired by CAM. Originally, CAM was introduced for object localization for CNNs, let $\boldsymbol{F} \in \mathbb{R}^{H \times W \times D}$ denote feature maps of the last convolutional layer with height H, width W and depth D, CAM $\boldsymbol{M}_k$ for class k reads [7]:

$$\boldsymbol{M}_k = \sum_{i=1}^{D} w_i^k \boldsymbol{F}_i, \tag{1}$$

where $\boldsymbol{F}_i$ and w_i^k represent the i-th channel of $\boldsymbol{F}$ and the corresponding weight in the fully connected layer for class k respectively. As CAM highlights the contribution location for each class, it enabled us to find out if the contextual information of the initially predicted label appears in the center area of the input patch. The calibration module defines a center-focusing area and calculates per-class score within this area for reselecting the correct class from the top-n predictions. Let φ represents the center-focusing area, $f_i(x, y)$ denotes values of $\boldsymbol{F}_i$ at location (x, y), where $(x, y) \in \varphi$. Considering $H \times W$ of $\boldsymbol{F}$ in most of modern CNNs is considerably small (e.g., 7×7), in order to define a relatively smaller focusing area φ, upsampling feature maps $\boldsymbol{F}$ is needed (e.g., upsample to the same size as input by bilinear interpolation). Score S for class k from area φ can be obtained:

$$S_k^\varphi = \sum_{x,y}^{\varphi} \sum_{i=1}^{D} w_i^k f_i(x, y). \tag{2}$$

If the output of softmax layer fulfils a criterion (which will be introduced later), scores for top-n ranked classes $\boldsymbol{S}^\varphi = [S_1^\varphi, S_2^\varphi, ..., S_n^\varphi]$ will be calculated using Eq. 2. The class with highest score will be accepted as the final prediction. Nevertheless, this calibration is only valid for the contextual patch contains more than one land cover classes. If the confidence score of the initial output is high, it is likely that the patch contains only one class, the proposed calibration is unnecessary for this scenario. In this work, we utilize the highest probability output by the softmax layer to define a criterion to determine if the proposed calibration is applicable for a given patch. Let $\boldsymbol{P} = [P_1, P_2, ..., P_C]$ denotes the probability vector in descending order for total C classes, if the highest probability $\boldsymbol{P}_1$ is lower than a threshold α, it can be determined that confidence of the prediction is insufficient and calculation for $\boldsymbol{S}^\varphi$ for calibrating the prediction should be applied. Compared with simply reducing the contextual area, the proposed method is designed to perform calibration when needed, meanwhile well maintain the general classification performance.

2.2. Shift-Invariance Improvement

The shift-invariance capability turns out to be poorly preserved by modern CNNs. In order to achieve high task performance, subsampling in modern CNNs (e.g., max-pooling) neglected an important sampling theorem: the signal should be blurred before subsampling, which results in the insufficient shift-invariance capability as suggested in [8, 9]. Meaning when sliding the window in boundary area at a small stride, the slight shift would lead to drastic deviations in feature maps in deep convolutional layers. This would eventually affect the function of calibration module because

Table 1: Numerical results on the test dataset. Per-class accuracy, OA and AA are displayed in percentage.

Network	Forest	Grass.	Water	Lar. B.	Sma. B.	Imp. S.	OA	AA	Kappa
ResNet-18	58.3	55.4	64.1	**81.0**	70.4	75.9	66.1	67.5	0.593
SICF-ResNet-18	**69.2**	**81.4**	**74.3**	75.5	**78.1**	**87.3**	**77.4**	**77.6**	**0.728**
DenseNet-121	51.8	60.1	61.3	**83.2**	74.6	72.6	65.9	67.3	0.590
SICF-DenseNet-121	**82.2**	**71.6**	**84.9**	71.7	**80.8**	**77.8**	**78.5**	**78.2**	**0.741**
MobileNetV2	37.3	65.5	**69.0**	**74.5**	58.1	**82.5**	63.2	64.5	0.559
SICF-MobileNetV2	**52.9**	**72.3**	42.6	72.8	**76.9**	80.2	**65.1**	**66.3**	**0.581**

feature maps in the last convolutional layer are used as input for $\boldsymbol{S}^{\varphi}$ calculation.

To our best knowledge, there are two effective solutions for improving the shift-invariance up to date. One possible approach is removing / reducing subsampling layers in CNN [9], however we believe this method would lead to high computational costs for land cover classification task at large scale. In this work, we employ the antialiasing solution provided in [8]. More detailedly, max-pooling is replaced by three operations: firstly perform max-pooling with stride 1, secondly apply low-pass filter, finally subsample. Similarly, we follow the suggestion from the researchers of this method, substitute non-strided convolutional layer followed by antialiasing and subsampling operations for each strided convolutional layer in the original network. In principle, the modifications should deliver a shift-invariant feature extraction system to pave the way for calibration module to function normally. Fig. 2 demonstrates the comprehensive proposed network.

3. EXPERIMENTS

3.1. Data Description

In this work, we acquire Google Earth zoom-level 16 optical image from Berlin area and resample the spatial resolution to 2 meters per pixel to generate a dataset containing six common land cover classes: forest, grassland, water, large built-up area, small built-up area and impervious surface. The size of the patch is 128 × 128 pixels. Depends on the way of contextual patch sampling, the dataset can be differentiated into two parts. The first part consists of $\sim$10500 patches, each patch encloses only one single land cover class contextual information. We divide the first part in the ratio of 4:1 for training and selecting the model. As the aim of this research is to deliver more accurate boundaries, the second part comprises $\sim$1500 patches are sampled from different boundary areas, meaning each of them contains contextual information correspond to more than one classes (i.e., more than one land cover classes present in the image). The second part is used as testing set to evaluate the classification performance.

In addition, we extract a region where a clear boundary separating two land cover classes presents as a test area to simulate the real land cover classification task to provide a qualitative evaluation.

3.2. Experiments Setup

We apply the proposed method to three modern CNNs to exam the performance: ResNet-18 [10], DenseNet-121 [11] and MobileNetV2 [12]. As mentioned previously, for each CNN, we modify the max-pooling and strided convolutional layer to build the shift-invariant feature extraction module and adhere the calibration module to the softmax layer to build the corresponding SICF implementation.

The input patch is upscaled to 224 × 224 pixels when fed into the network. The center-focusing area φ is defined as the innermost 16 × 16 pixels ($\boldsymbol{F}$ is bilinearly interpolated to the same size as input). We set the threshold α to 0.999, if the highest probability of the prediction is lower than this threshold, score $\boldsymbol{S}^{\varphi}$ for the top-3 classes will be calculated for calibration.

The experiments are conducted with PyTorch on an Nvidia 2080 GPU card, for comparison, the three employed CNNs and their corresponding SICF implementations are trained for 1k epochs with Adam optimizer, learning rate 1e-4, batch size 16 and multi-class cross-entropy loss function.

3.3. Results and Discussions

We use overall accuracy (OA), average accuracy (AA), Cohen's Kappa coefficient and per-class result as the evaluation metrics. Table 1 lists the numerical results on the test dataset. From the per-class perspective, the land cover classes with homogeneous contextual information (e.g., grassland, forest, small built-up area), the proposed SICF method outperforms all the competitors. Whereas for the class with inhomogeneous contextual pattern such as large built-up area, the SICF performance is inadequate, we believe this is because the large built-up area usually feature complex geometry, the main contribution of the score might not come from the center area confined by the innermost 16 × 16 pixels. Moreover, surprisingly, SICF fails to improve MobileNetV2 for classifying water and impervious surface, one possible reason could be the shift-invariant modification for MobileNetV2 impairs the feature extraction performance for certain classes. While

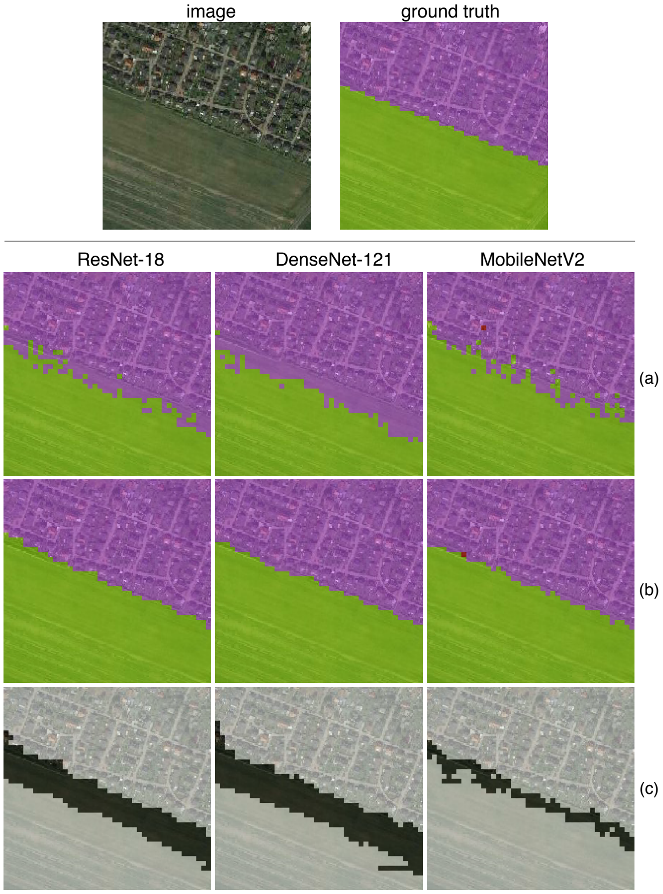

Fig. 3: Comparison of the land cover classification results from ResNet-18, DenseNet-121 and MobileNetV2 with (a) original architecture and (b) corresponding SICF network. Green and purple colours represent grassland and small built-up area respectively. The red pixels indicate where the patches are mis-classified as large built-up area. The dark areas in (c) indicate where the sampled patch meet the defined criterion for applying the calibration. It is worth mentioning that the stride for sliding window within the test area is 5, thus the size of the pixel in the ground truth and results is five times as large as the input.

from the global perspective, the numerical results clearly indicate that our proposed network surpasses all the competitors. As a dedicated solution for improving boundary area classification, the significantly enhanced performance is expected.

Fig. 3 illustrates the land cover classification results of the test area. The boundaries in all of the classification maps generated by the competing CNNs are conspicuously deviated from the one in ground truth, the discrepancies are unneglectable in a real land cover classification task. By contrast, SICF networks deliver highly accurate boundaries. Compared with the conventional CNNs, the advantages of our proposed method are significant.

4. CONCLUSIONS

In this work, we propose a shift-invariant center-focusing (SICF) network to deliver more accurate boundaries for the patch-based land cover classification. This network is designed to calculate the class score from a center-focusing area based on feature maps provided by a shift-invariant feature extraction system to calibrate the prediction. We perform evaluations quantitatively on a dataset sampled from boundary areas and visually on a test area, we can draw a conclusion that compared with the conventional CNNs, SICF network can significantly improve classification performance for contextual patches enclosing more than one land cover classes, hence delivers more accurate boundaries for land cover classification. Further researches including evaluating on more comprehensive land cover classes and discovering a more sophisticated definitiation of the center-focusing area for covering inhomogeneous contextual information are desired.

5. ACKNOWLEDGEMENTS

This work is jointly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*) and the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (Grant number: 01DD20001).

6. REFERENCES

[1] X. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing : A Comprehensive Review and List of Resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] J. Song, S. Gao, Y. Zhu, and C. Ma, "A Survey of Remote Sensing Image Classification Based on CNNs," *Big Earth Data*, vol. 3, no. 3, pp. 232–254, 2019.

[3] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks," *Remote Sensing*, vol. 8, no. 4, 2016.

[4] M. Mahdianpari, B. Salehi, M. Rezaee, F. Mohammadimanesh, and Y. Zhang, "Very Deep Convolutional Neural Networks for Complex Land Cover Mapping Using Multispectral Remote Sensing Imagery," *Remote Sensing*, vol. 10, no. 7, 2018.

[5] C. Yang, F. Rottensteiner, and C. Heipke, "Classification of Land Cover and Land Use Based on Convolutional Neural Networks," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 3, pp. 251–258, 2018.

[6] K. Zhou, D. Ming, X. Lv, J. Fang, and M. Wang, "CNN-based Land Cover Classification Combining Stratified Segmentation and Fusion of Point Cloud and Very High-Spatial Resolution Remote Sensing Image Data," *Remote Sensing*, vol. 11, no. 17, 2019.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *CVPR*, 2016.

[8] R. Zhang, "Making Convolutional Networks Shift-Invariant Again," in *ICML*, 2019.

[9] A. Azulay and Y. Weiss, "Why Do Deep Cnvolutional Networks Generalize So Poorly to Small Image Transformations?," *Journal of Machine Learning Research*, vol. 20, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*, 2017.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted Residuals and Linear Bottlenecks," in *CVPR*, 2018.