CrossRef DOI of original article:

Application of Convolutional Neural Network in the Segmentation and Classification of High-Resolution Remote Sensing Images

Dr. Ekambaram Kesavulu Reddy

Received: 1 January 1970 Accepted: 1 January 1970 Published: 1 January 1970

Abstract

Numerous convolution neural networks increase accuracy of classification for remote sensing scene images at the expense of the models' space and time sophistication. This causes the model to run slowly and prevents the realization of a trade-off among model accuracy and running time. The loss of deep characteristics as the network gets deeper makes it impossible to retrieve the key aspects with a sample double branching structure, which is bad for classifying remote sensing scene photos.

Index terms— remote sensing, convolutional neural network, standard convolution, feature extraction.

1 I. Introduction

n applications including urban development, land-use planning, infrastructure construction management, natural disasters, and crisis management, urban land-use classification is crucial [1]. The rate of change in land usage increases with the nation's rate of growth. Costly, labor-intensive, and time-consuming are landuse surveys [2]. China conducts a national land-use survey every ten years. High-resolution remote sensing processing technologies are being developed, which may assist planners in quickly and affordably gathering comprehensive land-cover data [3]. Deep convolutional neural networks (DCNNs), for instance, might fully achieve the classification of urban land-use by automatically extracting species-specific information from remote sensing photos. According to current criteria for land-use classification, one typical class may include more than one type of item. Each might also contain various objects that adhere to various standards. The Land-Use Standard of the 2nd and 3rd National Land-Use Resource Surveys, for instance, has various contents. Convolutional neural networks (CNN) trying to identify high-resolution remote sensing images are faced with significant difficulties by the complex spatial and textural patterns in one class [4]. Early FCNbased models had a limited ability to reconstruct spatial information despite acquiring rich contextual data and suffered from loss of high-frequency details, blurring boundaries, and difficulty identifying features.A skip connection was introduced to the networks to address this issue. By combining the multi-layer feature maps from the encoder with the decoder structure for incremental up sampling, Ronneberger et alU-Net.'s.

Architecture created high-resolution feature maps [5]. The classification effects of object boundaries are improved by the merging of high-and low-level semantic information. Later, Yu and Koltun added atrous convolution to fully convolutional networks (FCN), which could maintain the resolution of a featured image, expand the receptive field to capture multi-scale context information, and boost the precision of semantic segmentation using spatial information in the images [6]. Spatial Pyramid Pooling (SPP) [7] has been widely used to better capture information about the global context.

To take advantage of the potential of global context information, Zhao et al. used a pyramid pooling module to aggregate the context of several regions [8]. In order to gather multi-scale information, Chen et al. implemented pyramid-shaped atrous pooling in spatial dimensions [9] and piled up atrous convolution [10] with various atrous in cascade or in parallel. The resolution in the scale axis dimension was insufficient for Atrous Spatial Pyramid Pooling (ASPP) [9] to precisely extract target features from remote sensing images, therefore it still had certain drawbacks (RSIs). In order to effectively identify complicated situations while maintaining the model's size, Yang et al. introduced densely-connected Atrous Spatial Pyramid Pooling (DenseASPP) [12], which was able to cover a broader scale of the feature map and acquire more intensive receptive field information. A labor-intensive foundation, including welltagged remote sensing image labels for the most recent urban land cover types under distinct categorization standards, must be built in order to address the inherent difficulties in the present classification methods used to classify urban land use. Combining algorithms to produce higher-level sematic class images is another effective way to replace the original photos in laborintensive tasks. We proposed a double-layer deep convolutional neural network called DUA-Net that mainly combined two networks with different advantages, U-Net and DenseASPP, into a parallel structure in order to take into account the characteristics of urban land-use types, which contain multiple elements in one type. The methodology utilised in this study can produce a larger, continuous block of land use classification for urban areas. It can considerably cut down on operation durations and manual interactions when using the image of this classification result as the input for artificial fine classification, which can increase efficiency. Additionally, with the aid of vector data, we can fully utilise the same standard to categorise the photographs taken at various points in time in order to study the changes in land types at various times.

High resolution remote sensing images are being used in a variety of applications, including the classification of remote sensing scenes [1], hyperspectral image classification [2], change detection [3,4], geographic image classification, and the classification of land use [5, etc. However, image categorization presents significant challenges due to the intricate spatial patterns and geographical structure of remote sensing images. Therefore, it is crucial to effectively comprehend the semantic information of remote sensing photographs [6]. The goal of this study is to identify a straightforward and effective lightweight network model that is capable of quickly and reliably classifying remote sensing scene photos. Researchers have suggested a variety of techniques for efficiently extracting visual information.To begin with, manually created feature descriptors including colour histograms, texture descriptors, local binary mode, GIST, directional gradient histograms, bag-of-visual words (BOVW), etc. were used to extract picture features. Researchers then proposed some unsupervised feature learning techniques that can automatically extract shallow detail features from images in order to address the drawbacks of the method of manually extracting features. These techniques include principal component analysis (PCA), sparse coding, autoencoder, Latent Dirichlet allocation, and probabilistic latent semantic analysis. For the extraction of shallow picture information, the two feature extraction techniques mentioned above work quite well. However, the extraction of high-level features from images using these techniques is challenging, which restricts the development of classification accuracy.

Researchers have proposed convolutional neural networks, which have the ability to automatically extract significantly discriminative features from images, as a way to get around the limitations of existing methods. Since then, the model based on convolution neural networks has replaced other techniques as the industry standard for classifying remote sensing scene images. A lightweight convolution neural network may now achieve a balance between the speed of model operation and the precision of model classification thanks to advancements in convolution neural networks. Lightweight networks have currently been used for a variety of applications, including target recognition, image segmentation, and classification. The fire module, which separates the initial basic convolution layer into an extrusion layer and an expansion layer, was proposed by SqueezeNet. The extension layer is made up of a set of continuous 1 x 1 convolution and 3 x convolution channels, whereas the extruded layer is made up of a continuous set of 1 x 1 convolution channels. The Google team's MobileNet has three iterations: V1, V2, and V3. In order to divide the regular convolution into depthwise convolution and 1 x convolution, MobileNetV1 employs depthwise separable convolution. This significantly decreases the number of network parameters and, to some extent, increases accuracy. An inverse residual module and a linear bottleneck structure were presented by MobileNetV2. The convolution of 1 x 1 for ascending dimension, 3 x depthwise separable convolution for feature extraction, and 1 x 1 convolution for dimension reduction were all applied to this bottleneck structure in that order. With the addition of the SE module and the use of neural structure search, MobileNet V3 examines the network's setup and parameters [10]. An extremely effective convolution neural network architecture called ShuffleNet was created for mobile devices with constrained processing resources. Compared to some sophisticated ones with comparable accuracy, the design only requires two operations-group convolution and channel mixing-which significantly lowers the computation time.

2 II. Related Works

Remote sensing picture databases are being produced in greater numbers. These datasets use a variety of land cover and land use categories, and Castillo-Navarro et al. [14] have developed datasets that cover various scenes to increase surface coverage. Additionally, the labels that have been applied to the datasets vary [15]. For instance, BigEarthNet [17] and SEN12MS [16] both give image-level labels and pixellevel labels, respectively, and both datasets with varying scene categories can only be used for particular semantic segmentation applications. For instance, LULC has hundreds of fine-grained semantic categories like highways, buildings, cars, the countryside, cities, etc.

The circumstances that can show the relationship between the content of interest and its surroundings are rarely taken into account, and many datasets simply ignore the relationships within and across semantic classes [18]. Rich and detailed geometric features, texture information, and geospatial data are all present in high-resolution RSIs [19]. For land-use classification, the features extracted from these images can be interpreted with high accuracy. Pixelbased image analysis, object-based image analysis, and pixel-level semantic segmentation have all been used to classify the land use of RSIs [20]. Lowresolution remote sensing photos have historically been classified primarily using spectral data from remote sensing photographs.

Because the spectral features of pixels, which lack textural features and structural data, are unable to fully capture the characteristics of land-use kinds, the classification results for complicated land-use types, such as residential land and wasteland, are frequently less than optimal [21]. Similar pixels in different land-use types on residential and industrial land may exist. Some strategies, like Transfer Learning [22], Active Learning [24], and others, have been developed with the goal of increasing the size and enhancing the effectiveness of training datasets. Ammour et al. [25] merged two asymmetric networks for data domain adaption and classification, projected the two networks to the same feature space, and performed post-training for the weight coefficient adjustment method of the two networks. They employed a pretraining network for feature extraction.

Migration tests were conducted by Zhou et al. [26] using data from the same sensor at various dates. Additionally, they created a very difficult migration experiment that tested the efficacy of feature extraction and migration structure and was performed on hyperspectral remote sensing data from various viewpoints. The object-oriented classification approach [27] takes into account the correlation information between pixels and the internal texture features of ground objects while leveraging the spectrum information of RSIs [28] to make up for the inadequacies of conventional pixel-based classification methods. However, feature descriptions are generally incomplete, and the data collected is frequently insufficient to assist the characterization and identification of ground objects.

Deep learning overcomes the limitations of artificial features, directs object categorization, and achieves pixel-level land-use classification of RSIs by mastering the shape and texture aspects of various objects. Deep learning has been used extensively in RSIs for land-use classification. To automatically train the representative and discriminative features in a hierarchical way for land-use scene classification, deep filter banks were proposed to integrate multicolumn stacked denoising sparse autoencoders (SDSAE) with Fisher vectors (FV) [29]. A land-use classification framework for photographs (LUCFP) was presented by Xu et al., and it was effectively used to automate the verification of land surveys in China [30]. Adaptive hierarchical image segmentation improvement, multilevel extraction of features, and multiscale supervised deep -learning models were integrated to accurately produce detailed maps for disparate urban areas from the fusion of the UHSR ortho mosaic and digital elevation model, taking into account the highlevel details in an ultrahigh-spatial-resolution (UHSR) unmanned aerial vehicle (UAV) dataset (DSM). Excellent potential was shown by this framework for the thorough mapping of varied urban areas [31]. Another cuttingedge hybrid approach is multi-temporal relearning using convolutional long short-term memory (LSTM) models. It integrates post-classification relearning with locational semantic segmentation and is effective at categorising complex LULC maps with multitemporal VHR pictures [32].

3 III. Methodology a) Proposed Architecture

Figure 1 depicts the model's overall structure, which is broken down into nine sections. We suggest a feature extraction structure for the network's shallow layers in the first and second groups. The maximum pool layer is used for down sampling in the third group, where standard convolution and depth-wise separable convolution are combined. This reduces the spatial dimensions of the input images and lowers the danger of overfitting from irrelevant features. The majority of representative features from remote sensing image are extracted by groups 4 through 8. For the purpose of extracting richer feature information, Groups 4 through 7 use the proposed dual branch multi-level feature intense fusion method.

To extract deep-level features from Group 8, we sequentially applied 1 x 1 standard convolution, 3 x 3 standard convolution, and 3 x 3 depth wise separable convolution. The multilevel characteristics are fully exchanged and fused on the basis of double branch fusion, which not only increases classification accuracy but also significantly speeds up the network and achieves a balance between accuracy and speed. The number of convolution channels in Groups 5 and 8 is also increased to 256 and 512, respectively, in order to extract more features. The feature information generated by the final fusion is used to calculate the likelihood of each scene category, and Group 9 is used for classification. Each layer in deep feature extraction structures from Group 4 to Group 7 may fully extract the data of the current layer through three branches, including Identity, 1 x 1 standard convolution, and 3 x 3 depth wise separable convolution. Additionally, the shallow information loss caused by network deepening can be successfully avoided by merging the features retrieved by 1 x 1 standard convolution with each prior layer. Batch normalization (BN) can speed up training and use a greater learning rate while reducing the network's reliance on parameter initialization. Additionally, there are far less remote sensing photos available for training compared to the natural image data collection. In order to gather spatial information, boundary information, multi-scale contextual information, and global contextual information, our proposed model utilised parallel modules. As a result, it was able to reduce border ambiguity and class imbalance, address the inaccurate, fragmented single element classification in urban land-use semantic segmentation, and increase urban land-use classification accuracy. This section showed the DUA-suggested Net's architecture for classifying urban land use. The essential components of the suggested design, including the U-Net module, DenseASPP module, and Channel Attention Fusion module, were then thoroughly explained. In this study, U-Net and DenseASPP, two different DCNNs, were used to build the distributed system of DUA-Net, which fully utilised the various benefits of these two types of networks in the semantic segmentation of RSIs. The suggested framework has three components, as seen in Figure 1: a backbone network, a parallel extracting features module, and a feature fusion module. First, the VGG16 network is introduced as the foundation for extracting the features in U-Net and DenseASPP. Second, we use the U-Net module and DenseASPP module to simultaneously collect various semantic pieces of information due to the complexity of land-use type, structure, and geographic distribution of abnormality.

For more specifics, the DenseASPP module aggregates semantic information at various scales to capture multi-scale contextual information and global contextual information, and the U-Net module fuses high-level and low-level semantic information to improve the extraction of spatial and boundary information. Then, to address the issue of improper segmentation caused by comparable characteristics of similar categories, the feature maps output by the U-Net module and DenseASPP module were fused in the channel dimension through the attention mechanism in the Channel Attention Fusion module. The segmentation results were then produced by convolution with a convolution kernel size of 1 x 1 after the feature vectors had been mapped to the necessary number of classes. The DenseASPP module was introduced as the feature extractor in order to gather multi-scale contextual information as well as global contextual information in RSIs. In order to achieve integration at various levels with various dilation rates, DenseASPP employs the concept of dense connection and arranges numerous convolution layers in a cascade manner. Without significantly growing the model size, this organising approach not only covers a broader scale, but also covers it intensively. To gather semantic information from various scales, this work specifically exploited dense connections to send the output of each atrous convolution layer to all previously unvisited atrous convolution layers.

Additionally, the atrous convolution dilation rate at each layer increased layer by layer, enlarging the receptive field while maintaining the same level of feature map resolution. The layer with the lowest dilation rate among them was positioned in the lower layer, and the layer with the highest dilation rate was positioned in the upper layer. The outputted feature map from the multi-scale convolution process was the last step. The connection between feature channels is typically ignored by traditional techniques, and thus exhibit low sensitivity to critical information characteristics throughout the fusion process.

We used the channel attention method to successfully combine the feature maps from the U-Net module and the DenseASPP module. This fusion module achieved the automatic selection and weight assignment of attention regions, then increased output feature quality by using SENet to learn the correlation between various feature channels (and to boost the extraction of significant features). In particular, its primary operations were concatenation, squeeze, and excitation. The shallow features of the network are intended to be extracted by the first and second sets of down sampling structures. The impact of down sampling on network performance during the shallow feature extraction phase is significant. Down sampling is the process of scaling down the complex feature map to maintain the image's primary features while reducing the spatial size of the image. Having a greater number of pooling layers is one of the primary techniques for down sampling in deep convolution neural networks.

4 IV. Results and Discussion

Three of the down sampling techniques indicated in Section IIB are utilised in the first and second levels of the network to verify the effectiveness of our suggested down sampling techniques. Two datasets, UC and RSSCN, were employed for the experiments, and the OA and Kappa were used as evaluation metrics. According to Figure 2, the first and third convolution steps for the Conv-Downsampling (CD) are 1, while the second and fourth convolution steps are 2. The convolution kernels for the pooling down sampling (Max pooling-Downsampling, MD) are all 3 x 3, with convolution steps of 1 x 1. The pooling step size and maximum pooling size are both 2.On the two datasets, pooling down sampling had poorer classification accuracy and Kappa values than convolution down sampling. Convolution down sampling in deep networks produces superior nonlinear performance than pooled down sampling, which is the reason. On the 80/20UC and 50/50RSSCN datasets, the suggested down sampling methods have classification accuracy scores of 99.53% and 97.86%, respectively, and Kappa values of 99.50% and 97.50%, which are greater than those of the other two down sampling methods.

This demonstrates once again how much more accurately the multi-level features dense fusion technique can identify remote sensing scene photos. In this section, three types of visualisation, including grad cam, t-distribution random neighbour embedding (T-SNE), and randomly picked and tested are explained and examined in order to more clearly demonstrate the effectiveness of the suggested method. Through a visual thermal map, the grad cam presents the retrieved features in order of significance. The most comprehensive spatial and semantic information is found in the final layer of a convolution neural network.

Grad Cam creates an attention map to highlight key portions of an image by fully utilizing the features of the last layer of convolution. In this experiment, some remote sensing scene photos from the RSSCN collection of "Industries," "Fields," "Residences," "Grass," and "Forests" are randomly chosen. Figure 2 compares the thermal diagram visualization outcomes of the enhanced BMDF-LCNN approach with the baseline LCNN-BFF method.

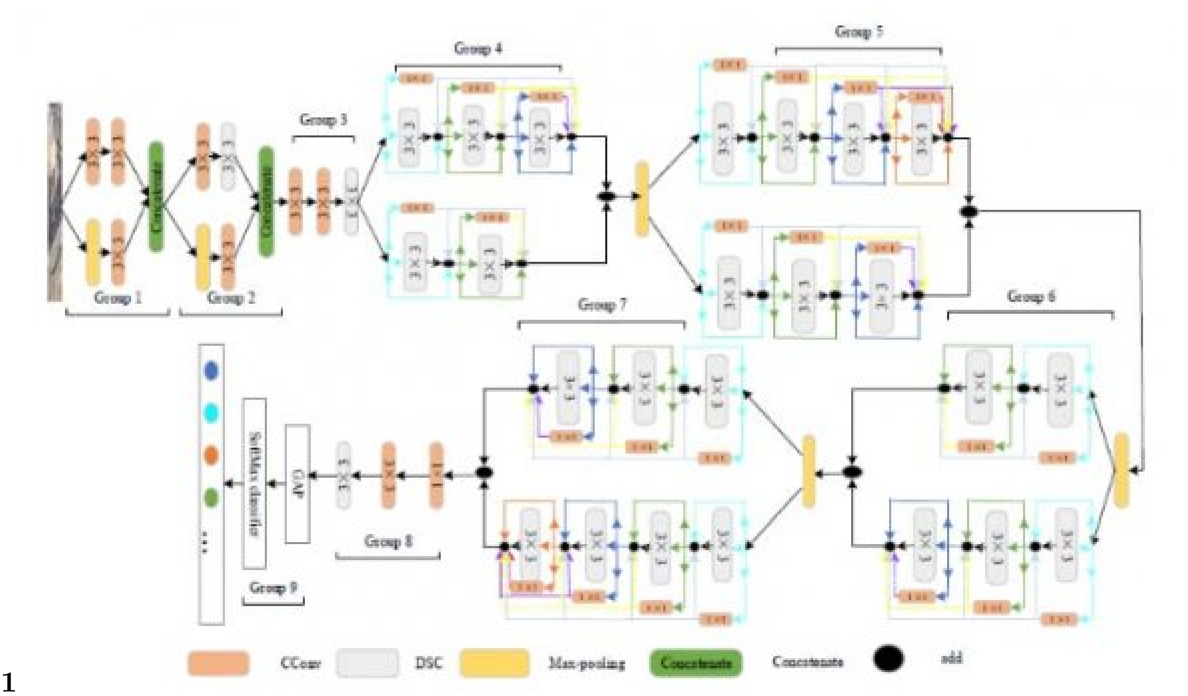
Figure 2 shows that, for "Industries" scenarios, the LCNN-BFF approach transfers the attention to the highway rather than accurately focusing on the factory region, whereas the proposed BMDF-LCNN method accurately focuses on the industrial area. While the BMDF-LCNN approach is well focused on the target region, the LCNN-BFF model's focused areas for the "Fields" and "Grass" scenarios both showed a partial deviation, ignoring the similar surrounding targets and searching with few objects. Additionally, the LCNN-BFF method's restricted coverage and inability to fully extract the target for scenario regions like "Residence" and "Forests" has an impact on the classification accuracy. However, in these cases, the suggested BMDF-LCNN approach can get a comprehensive region of interest. Next, we use t-distribution random neighbor embedding to illustrate the classification results on the UC (8/2) and RSSCN

(5/5) datasets (T-SNE).

High-latitude characteristics are mapped by T-SNE to two-or threedimensional space for visualization, which is a very effective way to assess the classification effect of the model.

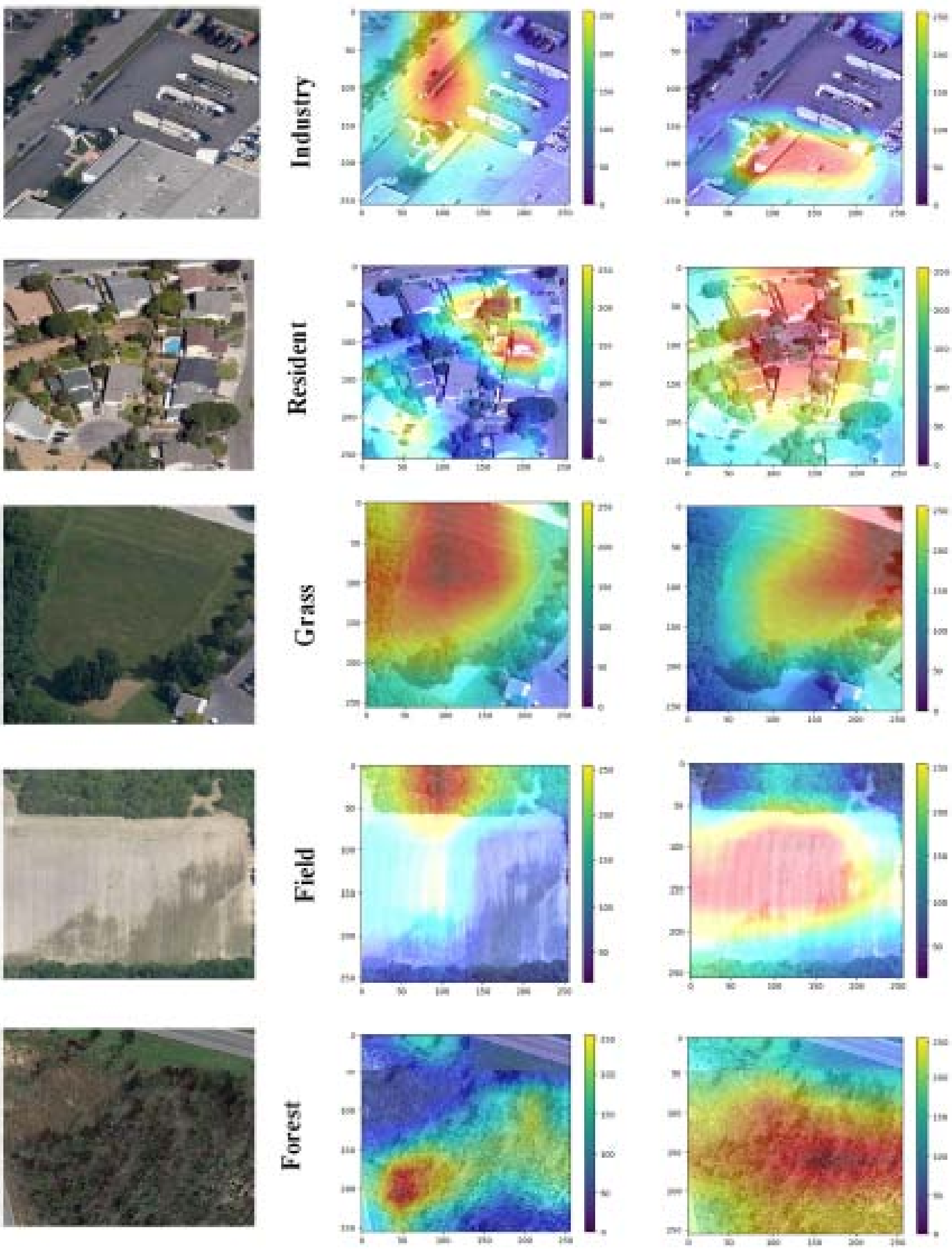
5 V. Conclusion

A lightweight network based on the dense fusion of dual-branch, multi-level features is proposed for the categorization of remote sensing scene photos. A fresh down sampling technique was also developed to gather more accurate feature data. The information of the current layer can be fully extracted and fused with the features extracted by 1x1 standard convolution in the previous layer using the three branches of 3 3 depth wise separable convolution, 1 x 1 standard convolution, and identity in the network. This effectively realizes the information interaction between different levels of features and improves the classification performance and computational speed of the model. The suggested model still requires development. Due to the generation of certain redundant data during multi-level feature heavy fusion, the computational complexity rises. Future



1

Figure 1: Figure 1 :



2

Figure 2: Figure 2 :

[Li et al. ()] 'A deep learning method for change detection in synthetic aperture radar images'. Y Li , C Peng , Y Chen , L Jiao , L Zhou , R Shang . *IEEE Trans. Geosci. Remote Sens* 2019. 57 p. .

[Xu et al. ()] 'A Framework for Land Use Scenes Classification Based on Landscape Photos'. S Xu , S Zhang , J Zeng , T Li , Q Guo , S Jin . *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens* 2020. 13 p. .

[Chen et al. ()] 'A Review of Fine-Scale Land Use and Land Cover Classification in Open-Pit Mining Areas by Remote Sensing Techniques'. W Chen , X Li , H He , L Wang . *Remote Sens* 2018. 10 p. 15.

[Liu et al. ()] 'Active Deep Learning for Classification of Hyperspectral Images'. P Liu , H Zhang , K B Eom . *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens* 2017. 10 p. .

[Ammour et al. ()] 'Asymmetric Adaptation of Deep Features for Cross-Domain Classification in Remote Sensing Imagery'. N Ammour , L Bashmal , Y Bazi , M M Al Rahhal , M Zuair . *IEEE Geosci. Remote Sens. Lett* 2018. 15 p. .

[Liu et al.] *Bidirectionalconvolutional LSTM based spectral-spatial feature learning for hyperspectral image classifification*, Q Liu , F Zhou , R Hang , X Yuan . p. 1330. (Remote Sens. 2017, 9)

[Sumbul et al. (2019)] 'Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding'. G Sumbul , M Charfuelan , B Demir , V Markl . *Proceedings of the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, (the IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing SymposiumYokohama, Japan) 28 July-2 August 2019. p. .

[Swain and Ballard ()] 'Color indexing'. M J Swain , D H Ballard . *Int. J. Comput. Vis* 1991. 7 p. .

[Wilhelm et al. (2021)] 'Cover Classification from a Mapping Perspective: Pixelwise Supervision in the Deep Learning Era'. T Wilhelm , D Koßmann , Land . *Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, (the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSSBrussels, Belgium) July 2021. p. .

[Zhou and Prasad ()] 'Deep Feature Alignment Neural Networks for Domain Adaptation of Hyperspectral Data'. X Zhou , S Prasad . *IEEE Trans. Geosci. Remote Sens* 2018. 56 p. .

[Wu et al. ()] 'Deep Filter Banks for Land-Use Scene Classification'. H Wu , B Liu , W Su , W Zhang , J Sun . *IEEE Geosci. Remote Sens. Lett* 2016. 13 p. .

[Ma et al. ()] 'Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review'. L Ma , Y Liu , X Zhang , Y Ye , G Yin , B A Johnson . *ISPRS J. Photogramm. Remote Sens* 2019. 152 p. .

[Neupane et al.] *Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis*, B Neupane , T Horanont , J Aryal . p. 808. (Remote Sens. 2021, 13)

[Peng et al. ()] 'Densely based multiscale and multi-modal fully convolutional networks for highresolution remotesensing image semantic segmentation'. C Peng , Y Li , L Jiao , Y Chen , R Shang . *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens* 2019. 12 p. .

[Pan and Zhao ()] 'High-Resolution Remote Sensing Image Classification Method Based on Convolutional Neural Network and Restricted Conditional Random Field'. X Pan , J Zhao . *Remote Sens* 2018. 10 p. 920.

[Dalal and Triggs (2015)] 'Histograms of oriented gradients for human detection'. N Dalal , B Triggs . *Proceedings of the CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (the CVPR, IEEE Computer Society Conference on Computer Vision and Pattern RecognitionBoston, MA, USA) June 2015. p. .

[Lu et al. ()] 'Joint dictionary learning for multispectral change detection'. X Lu , Y Yuan , X Zheng . *IEEE Trans. Cybern* 2017. 47 p. .

[Lyu et al. ()] 'Long-Term Annual Mapping of Four Cities on Different Continents by Applying a Deep Information Learning Method to Landsat Data'. H Lyu , H Lu , L Mou , W Li , J Wright , X Li , X Li , X Zhu , J Wang , L Yu . *Remote Sens* 2018. 10 p. 471.

[Gibril et al.] *Mapping Heterogeneous Urban Landscapes from the Fusion of Digital Surface Model and Unmanned Aerial Vehicle-Based Images Using Adaptive Multiscale Image Segmentation and Classification*, M B A Gibril , B Kalantar , R Al-Ruzouq , N Ueda , V Saeidi , A Shanableh , S Mansor , H Z Shafri . p. 1081. (Remote Sens. 2020, 12)

[Oliva and Antonio ()] 'Modeling the shape of the scene: A holistic representation of the spatial envelope'. A Oliva , T Antonio . *Int. J. Comput. Vis* 2001. 42 p. .

[Wang et al. ()] 'Multiple resolution block feature for remote-sensing scene classification'. C Wang , W Lin , P Tang . *Int. J. Remote Sens* 2019. 40 p. .

[Ojala et al. ()] 'Multiresolution gray-scale and rotation invariant texture classifification with local binary patterns'. T Ojala , M Pietikainen , T Maenpaa . *IEEE Trans. Pattern Anal. Mach. Intell* 2002. 24 p. .

[Zhu et al. ()] 'Multitemporal Relearning with Convolutional LSTM Models for Land Use Classification'. Y Zhu , C Geis , E So , Y Jin . *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens* 2021. 14 p. .

[Ghamisi et al. ()] 'New frontiers in spectral-spatial hyperspectral image classifification: The latest advances based on mathematical morphology, Markov random fifields, segmentation, sparse representation, and deep learning'. P Ghamisi , E Maggiori , S Li , R Souza , Y Tarablaka , G Moser , Y Chen . *IEEE Geosci. Remote Sens. Mag* 2018. 6 p. .

[Walter ()] 'Object-Based Classification of Remote Sensing Data for Change Detection'. V Walter . *ISPRS J. Photogramm. Remote Sens* 2004. 58 p. .

[Long et al. ()] 'On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID'. Y Long , G S Xia , S Li , W Yang , M Y Yang , X X Zhu , L Zhang , D Li . *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens* 2021. 14 p. .

[Richards and Jia ()] *Remote Sensing Digital Image Analysis: An Introduction*, J A Richards , X Jia . 2005. Berlin, Germany: Springer. (4th ed.)

[Zhou et al. ()] 'Remote sensing scene classification based on rotation invariant feature learning and joint decision making'. Y Zhou , X Liu , J Zhao , D Ma , R Yao , B Liu , Y Zheng . *EURASIP J. Image* Video Process. 2019, 2019. p. .

[Song et al. (2010)] 'Rotation invariant texture measured by local binary pattern for remote sensing image classifification'. C Song , F Yang , P Li . *Proceedings of the 2nd International Workshop on Education Technology and Computer Science*, (the 2nd International Workshop on Education Technology and Computer ScienceWuhan, China, 6-7) March 2010. 3 p. .

[Schmitt et al. ()] *SEN12MS-A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion*, M Schmitt , L H Hughes , C Qiu , X X Zhu . arXiv:1906.07789. 2019.

[Hu et al. ()] *Transferring deep convolutional neural networks for the scene classifification of high-resolution remote sensing imagery. Remote Sens*, F Hu , G S Xia , J Hu , L Zhang . 2015. 7 p. .

[Sivic and Zisserman (2003)] 'Video Google: A text retrieval approach to object matching in videos'. J Sivic , A Zisserman . *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (the IEEE International Conference on Computer Vision (ICCV)Nice, France) October 2003. p. 1470.