(ABNet) Adaptive Balanced Network For Multiscale Object Detection in Remote Sensing Imagery

[1]MRS. RAVALI, [2]P. PRANATHI

[1](Assistant Professor),**ECE,** Sreyas Institute Of Engineering And Technology

[2]B.Tech scholar ,**ECE,** Sreyas Institute Of Engineering And Technology

ABSTRACT

Benefiting from the development of convolutional neural networks (CNNs), many excellent algorithms for object detection have been presented. Remote sensing object detection (RSOD) is a challenging task mainly due to: 1) complicated background of remote sensing images (RSIs) and 2) extremely imbalanced scale and sparsity distribution of remote sensing objects. Existing methods cannot effectively solve these problems with excellent detection accuracy and rapid speed. To address these issues, we propose an adaptive balanced network (ABNet) in this article. First, we design an enhanced effective channel attention (EECA) mechanism to improve the feature representation ability of the backbone, which can alleviate the obstacles of complex background on foreground objects. Then, to combine multiscale features adaptively in different channels and spatial positions, an adaptive feature pyramid network (AFPN) is designed to capture more discriminative features. Furthermore, considering that the original FPN ignores rich deeplevel features, a context enhancement module (CEM) is proposed to exploit abundant semantic information for multiscale object detection. Experimental results on three public datasets demonstrate that our approach exhibits superior performance over baseline by only introducing less than 1.5M extra parameters.

Keywords: Multiscale Object Detection, Remote Sensing Imagery, Adaptive Balanced Network, Enhanced Effective Channel Attention.

1. INTRODUCTION

With the development of aerial technology, the acquisitions and applications of remote sensing images (RSIs) have become more diverse. Remote sensing object detection (RSOD) is one of the hot research topics in the field of RSIs analysis. It not only locates the object regions of interest in RSIs but also categorizes the classes of multi objects, which has been widely used in hazard response, urban monitoring, traffic control, and so on. Although many algorithms have been proposed for RSOD, especially for largescale RSIs, this task still remains challenges mainly due to complex scenes and multiscale objects. Different from natural scene images, RSIs are commonly captured from satellites with wide views, which leads to the large-scale images and background clutter. Furthermore, objects in different RSIs are in various scales by reason of the variation in image acquisition altitudes. Besides, certain categories of objects are usually distributed densely in RSIs, such as ships and vehicles.

The above issues are the main obstacles for object detection in RSIs, which makes most algorithms for natural images not adapted to RSIs well. Most RSOD algorithms based on convolutional neural networks (CNNs) are motivated by corresponding methods for natural images. The mainstream object detection approaches can be roughly divided into two types: two-stage and one stage. The former defines the task as a step-bystep refining process (regions extraction and bounding boxes classification), while the latter performs a one-step process. 2 Faster RCNN is a representative two-stage method that implements the first end-to-end network for general object detection. Its main innovation is to design a region proposal network (RPN) to gather proposals instead of a sliding window. The typical one-stage methods mainly include YOLO and so on. For example, YOLO applies a single network to the input image and divides the image into several cells. Then, it outputs the predicted bounding boxes (b-boxes) and categories probabilities of each region directly. However, these algorithms are not good at dealing with multiscale objects.

• Current methods face challenges due to the diverse scales and conditions present in such imagery, as well as imbalances in the types of objects detected.

• By leveraging advanced techniques in deep learning and addressing these challenges, the project seeks to create a system that can efficiently adapt to different scales, handle

imbalanced data, and provide reliable object detection results.

• This system would not only improve the accuracy of detection but also streamline workflows and support decision-making processes in real-world scenarios

2. LITERATURE SURVEY

2.1 Embedding structured contour and location prior in Siamese fully convolutional networks for road detection: https://arxiv.org/abs/1905.01575

ABSTRACT:

Road detection from the perspective of moving vehicles is a challenging issue in autonomous driving. Recently, many deep learning methods spring up for this task because they can extract high-level local features to find road regions from raw RGB data, such as Convolutional Neural Networks (CNN) and Fully Convolutional Networks (FCN). However, how to detect the boundary of road accurately is still an intractable problem. In this paper, we propose a Siamese fully convolutional networks (named as "s-FCN-loc"), which is able to consider RGB-channel images, semantic contours and location priors simultaneously to segment road region elaborately. To be specific, the s-FCN-loc has two streams to process the original RGB images and contour maps respectively. At the same time, the location prior is directly appended to the Siamese FCN to promote the final detection performance. Our contributions are: (1) An s-FCN-loc is proposed that learns more discriminative features of road boundaries than the original FCN to detect more accurate road regions; (2) Location prior is viewed as a type of feature map and directly appended to the final feature map in s-FCN-loc to promote the detection performance effectively, which is easier than other traditional methods, namely different priors for different inputs (image patches); 4 (3) The convergent speed of training s-FCN-loc model is 30\% faster than the original FCN, because of the guidance of highly structured contours. The proposed approach is evaluated on KITTI Road Detection Benchmark and One-Class Road Detection Dataset, and achieves a competitive result with state of the arts.

2.2 Dual feature extraction network for hyperspectral image analysis:

https://www.researchgate.net/publication/351408907_Dual_Feature_Extraction_Network_for_Hyperspectral_Image_Analysis

ABSTRACT: Hyperspectral anomaly detection (HAD) is a research endeavor of

high practical relevance within remote sensing scene interpretation. In this work, we propose an unsupervised approach, dual feature extraction network (DFEN) for HAD, to gradually build up evergreater discrimination between the original data and background. In particular, we impose an end-to-end discriminative learning loss on two networks. Among them, adversarial learning aims to keep the original spectrum while Gaussian constrained learning intends to learn the background distribution in the potential space. To extract the anomaly, we calculate spatial and spectral anomaly scores based on mean squared error (MSE) spatial distance and orthogonal projection divergence (OPD) spectral distance between two latent feature matrices. Finally, the comprehensive detection result is obtained by a simple dot product between two domains to further reduce the false alarm rate. Experiments have been conducted on eight real hyperspectral data sets captured by different sensors over different scenes, which show that the proposed DFEN method is superior to other compared methods in detection accuracy or false alarm rate. 5

2.3 Hyperspectral pan sharpening with deep priors:

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8750899

ABSTRACT: Hyperspectral (HS) image can describe subtle differences in the spectral signatures of materials, but it has low spatial resolution limited by the existing technical and budget constraints. In this paper, we propose a promising HS pan sharpening method with deep priors (HPDP) to fuse a low-resolution (LR) HS image with a high-resolution (HR) panchromatic (PAN) image. Different from the existing methods, we redefine the spectral response function (SRF) based on the larger eigenvalue of structure tensor (ST) matrix for the first time that is more in line with the characteristics of HS imaging. Then, we introduce HF Net to capture deep residual mapping of high frequency across the up sampled HS image and the PAN image in a band by-band manner. Specifically, the learned residual mapping of high frequency is injected into the structural transformed HS images, which are the extracted deep priors served as additional constraint in a Sylvester equation to estimate the final HR HS image. Comparative analyses validate that the proposed HPDP method presents the superior pan sharpening performance by ensuring higher quality both in spatial and spectral domains for all types of data sets. In addition,

the HF Net is trained in the high-frequency domain based on multispectral (MS) images, which overcomes the sensitivity of deep neural network (DNN)to data sets acquired by different sensors and the difficulty of insufficient training samples for HS pan sharpening.

2.4 Weakly supervised low-rank representation for hyperspectral anomaly detection:
https://ieeexplore.ieee.org/document/9426593

ABSTRACT: In this article, we propose a weakly supervised low-rank representation (WSLRR) method for hyperspectral anomaly detection (HAD), which formulates deep learning-based HAD into a low-lank optimization problem not only characterizing the complex and diverse background in real HSIs but also obtaining relatively strong supervision information. Different from the existing unsupervised and supervised methods, we first model the background in a weakly supervised manner, which achieves better performance without prior information and is not restrained by richly correct annotation. Considering reconstruction biases introduced by the weakly supervised estimation, LRR is an effective method for further exploring the intricate background structures. Instead of directly applying the conventional LRR approaches, a dictionary-based LRR, including both observed training data and hidden learned data drawn by the background estimation model, is proposed. Finally, the derived low-rank part and sparse part and the result of the initial detection work together to achieve anomaly detection. Comparative analyses validate that the proposed WSLRR method presents superior detection performance compared with the state-of-the-art methods.

2.5 A joint convolutional neural networks and context transfer for street scenes labeling:
https://arxiv.org/pdf/1905.01574.pdf

ABSTRACT: Street scene understanding is an essential task for autonomous driving. One important step towards this direction is scene labeling, which annotates each pixel in the images with a correct class label. Although many approaches have been developed, there are still some weak points. Firstly, many methods are based on the hand-crafted features whose image representation ability is limited. Secondly, they cannot label foreground objects accurately due to the dataset bias. Thirdly, in the refinement stage, the traditional Markov Random Filed (MRF) inference is prone to over smoothness. For

improving the above problems, this paper proposes a joint method of priori convolutional neural networks at super pixel level (called as "priori s-CNNs") and soft restricted context transfer. Our contributions are threefold: (1) A priori s-CNNs model that learns priori location information at super pixel level is proposed to describe various objects discriminatingly (2) A hierarchical data augmentation method is presented to alleviate dataset bias in the priori s-CNNs training stage, which improves foreground objects labeling significantly (3) A soft restricted MRF energy function is defined to improve the priori s-CNNs model's labeling performance and reduce the over smoothness at the same time. The proposed approach is verified on Cam Vid dataset (11 classes) and SIFT Flow Street dataset (16 classes) and achieves a competitive performance

3.NEURAL NETWORKS

Ab net uses deep learning techniques such as:

A) ARTIFICIAL NEURAL NETWORK

B) CONVOLUTIONAL NEURAL NETWORKS BIOLOGICAL NEURAL NETWORK

Humans have made several attempts to mimic the biological systems, and one of them is artificial neural networks inspired by the biological neural networks in living organisms. However, they are very much different in several ways. For example, the birds had inspired humans to create airplanes, and the four-legged animals inspired us to develop cars. The artificial counterparts are definitely more powerful and make our life better.

The perceptron, who are the predecessors of artificial neurons, were created to mimic certain parts of a biological neuron such as dendrite, axon, and cell body using mathematical models, electronics, and whatever limited information we have of biological neural networks. For creating mathematical models for artificial neural networks, theoretical analysis of biological neural networks is essential as they have a very close relationship. And this understanding of the brain's neural networks has opened horizons for the development of artificial neural network systems and adaptive systems designed to learn and adapt to the situations and inputs. • Biological neurons provide the foundation for artificial neural networks in deep learning. These neurons, found in the human brain, communicate and process information through networks. 17 • Artificial neurons mimic the structure and function of biological ones, enabling computers to learn from data.

Like their biological counterparts, artificial neural networks adapt and improve through training. • Leveraging principles from biological neurons enhances the efficiency and complexity of deep learning algorithms.

A) ARTIFICIAL NEURAL NETWORK

Artificial neural networks (ANNs) are computational models inspired by the human brain's neural networks. Artificial Neural Networks contain artificial neurons which are called units. These units are arranged in a series of layers that together constitute the whole Artificial Neural Network in a system. A layer can have only a dozen units or millions of units as this depends on how the complex neural networks will be required to learn the hidden patterns in the dataset. Commonly, Artificial Neural Network has an input layer, an output layer as well as hidden layers. The input layer receives data from the outside world which the neural network needs to analyze or learn about. Then this data passes through one or multiple hidden layers that transform the input into data that is valuable for the output layer. Finally, the output layer provides an output in the form of a response of the Artificial Neural Networks to input data provided. In the majority of neural networks, units are interconnected from one layer to another. Each of these connections has weights that determine the influence of one unit on another unit. As the data transfers from one unit to another, the neural network learns more and more about the data which eventually results in an output from the output layer

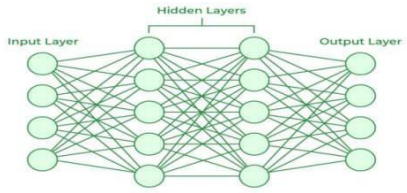

Fig 1: Neural Network Architecture

The structures and operations of human neurons serve as the basis for artificial neural networks. It is also known as neural networks or neural nets. The input layer of an artificial neural network is the first layer, and it receives input from external sources and releases it to the hidden layer, which is the second layer. In the hidden layer, each neuron receives input from the previous layer neurons, computes the weighted sum, and sends it to the neurons in the next layer. These connections are weighted means effects of the inputs from the previous layer are optimized more or less by assigning different-different weights to each input and it is adjusted during the training process by optimizing these weights for improved model performance. Artificial neural networks are

trained using a training set. For example, suppose you want to teach an ANN to recognize a cat.

Then it is shown thousands of different images of cats so that the network can learn to identify a cat. Once the neural network has been trained enough using images of cats, then you need to check if it can identify cat images 19 correctly. This is done by making the ANN classify the images it is provided by deciding whether they are cat images or not. The output obtained by the ANN is corroborated by a human-provided description of whether the image is a cat image or not. If the ANN identifies incorrectly then backpropagation is used to adjust whatever it has learned during training. Backpropagation is done by fine-tuning the weights of the connections in ANN units based on the error rate obtained. This process continues until the artificial neural network can correctly recognize a cat in an image with minimal possible error rates.

Applications of Artificial Neural Networks: Social Media: Artificial Neural Networks are used heavily in Social

• Media. For example, let's take the 'People you may know' feature on Facebook that suggests people that you might know in real life so that you can send them friend requests. Well, this magical effect is achieved by using Artificial Neural Networks that analyze your profile, your interests, your current friends, and also their friends and various other factors to calculate the people you might potentially know. Another common application of Machine Learning in social media is facial recognition. This is done by finding around 100 reference points on the person's face and then matching them with those already available in the database using convolutional neural networks. Marketing and Sales: When you log onto E-commerce sites like

• Amazon and Flipkart, they will recommend your products to buy based on your previous browsing history.

Similarly, suppose you love Pasta, then Zomato, Swiggy, etc. will show you restaurant recommendations based on your tastes and previous order history. This is true across all new-age marketing segments like Book sites, Movie services, Hospitality sites, etc. and it is done by implementing 20 personalized marketing. This uses Artificial Neural Networks to identify the customer likes, dislikes, previous shopping history, etc., and then tailor the marketing campaigns accordingly. Healthcare: Artificial Neural Networks are used in Oncology to train

• Algorithms that can identify cancerous tissue at the microscopic level at the same accuracy as trained physicians. Various rare diseases may manifest in physical characteristics and can be identified in their premature stages by using Facial Analysis on the patient photos. So the full-scale implementation of Artificial Neural Networks in the healthcare environment can only enhance the diagnostic abilities of medical experts and ultimately lead to the overall improvement in the quality of medical care all over the world. Personal Assistants: We all have heard of Siri, Alexa, Cortana, etc.

,• And also heard them based on the phones you have!!!

These are personal assistants and an example of speech recognition that uses Natural Language Processing to interact with the users and formulate a response accordingly. Natural Language Processing uses artificial neural networks that are made to handle many tasks of these personal assistants such as managing the language syntax, semantics, correct speech, the conversation that is going on, etc. Types of artificial neural networks: ANNs have evolved into a broad family of techniques that have advanced the state of the art across multiple domains. The simplest types have one or more static components, including number of units, number of layers, unit weights and topology. Dynamic types allow one or more of these to evolve via learning. The latter is much more complicated but can shorten learning periods and produce better results. Some types allow/require learning to be "supervised" by the operator, while others operate independently. Some types operate purely 21 in hardware, while others are purely software and run on general purpose computers. Some of the main breakthroughs include: Convolutional neural networks that have proven particularly

• Successful in processing visual and other two-dimensional data where long short-term memory avoids the vanishing gradient problem and can handle signals that have a mix of low and high frequency components aiding large-vocabulary speech recognition, text-to-speech synthesis and photo-real talking heads. Competitive networks such as generative adversarial networks in

• Which multiple networks (of varying structure) compete with each other, on tasks such as winning a game or on deceiving the opponent about the authenticity of an input. Using artificial neural networks requires an understanding of their characteristics. Choice of model:

• This depends on the data representation and the application. Model parameters include the number, type, and connectedness of network layers, as well as the size of each and the connection type (full, pooling, etc.). Overly complex models learn slowly. Learning algorithm

• Numerous trade-offs exist between learning algorithms. Almost any algorithm will work well with the correct hyperparameters] for training on a particular data set. However, selecting and tuning an algorithm for training on unseen data requires significant experimentation. Robustness: If the model, cost function and learning algorithm

• Are selected appropriately, the resulting ANN can become robust.

ADVANTAGES OF ARTIFICIAL NEURAL NETWORK:

A neural network can implement tasks that a linear program cannot. When an item of the neural network declines, it can continue without some issues by its parallel features. A neural network determines and does not require to be reprogrammed .It can be executed in any application.

DISADVANTAGES OF ARTIFICIAL NEURAL NETWORK:

The neural network required training to operate.

• The structure of a neural network is disparate from the structure of microprocessors therefore required to be emulated. It needed high processing time for big neural networks.

B) CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Networks (CNNs) are a type of deep neural network specifically designed for processing structured grid data, such as images. A convolutional neural network (CNN) is a category of machine learning model, namely a type of deep learning algorithm well suited to analyzing visual data. CNNs -- sometimes referred to as convnets -- use principles from linear algebra, particularly convolution operations, to extract features and identify patterns within images. Although CNNs are predominantly used to process images, they can also be adapted to work with audio and other signal data. CNN architecture is inspired by the connectivity patterns of the human brain -- in particular, the visual cortex, which plays an essential role in perceiving and processing visual stimuli.

The artificial neurons in a CNN are arranged to efficiently interpret visual information, enabling 23 these models to process entire images. Because CNNs are so effective at identifying objects, they are frequently used for computer vision tasks such as image recognition and object detection, with common use cases including self-driving cars, facial recognition and medical image analysis. Unlike CNNs, older forms of neural networks often needed to process visual data in a piecemeal manner, using segmented or lower- resolution input images.

A CNN's comprehensive approach to image recognition lets it outperform a traditional neural network on a range of image-related tasks and, to a lesser extent, speech and audio processing. CNNs use a series of layers, each of which detects different features of an input image. Depending on the complexity of its intended purpose, a CNN can contain dozens, hundreds or even thousands of layers, each building on the outputs of previous layers to recognize detailed patterns. The process starts by sliding a filter designed to detect certain features over the input image, a process known as the convolution operation (hence the name "convolutional neural network"). The result of this process is a feature map that highlights the presence of the detected features in the image. This feature map then serves as input for the next layer, enabling a CNN to gradually build a hierarchical representation of the image. Initial filters usually detect basic features, such as lines or simple textures. Subsequent layers' filters are more complex, combining the basic features identified earlier on to recognize more complex patterns. For example, after an initial layer detects the presence of edges, a deeper layer could use that information to start identifying shapes. 24 Between these layers, the network takes steps to reduce the spatial dimensions of the feature maps to improve efficiency and accuracy. In the final layers of a CNN, the model makes a final decision -- for example, classifying an object in an image -- based on the output from the previous layers. Before we go to the working of Convolutional neural networks (CNN), let’s cover the basics, such as what an image is and how it is represented. An RGB image is nothing but a matrix of pixel values having three planes whereas a grayscale image is the same but it has a single plane.

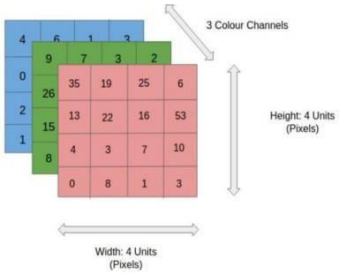

Fig 2: Matrix Of an Image

WORKING OF CONVOLUTIONAL NEURAL NETWORK:

CNN typically consists of several layers, which can be broadly categorized into three groups: convolutional layers, pooling layers and fully connected layers. As data passes through these layers, the complexity of the CNN increases, which lets the CNN successively identify larger portions of an image and more abstract features. Convolutional Neural Networks (CNNs) typically consist of several types of layers, each serving a specific purpose in extracting and processing features from input data.

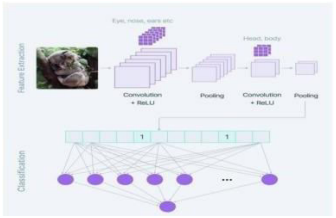

Fig 3: Convolutional Neural Network Architecture

4. OBSERVATIONS

INPUT AND OUTPUT

YOLO takes an input image containing various objects, including an airplane.

Fig 4: Input Image

The input image taken is preprocessed to prepare it for input into the neural network.

•This involve resizing the image to a standard size, normalizing pixel values, and other transformations. The preprocessed image is passed through the YOLO neural network.

• It consists of convolutional layers, pooling layers, and other specialized layers. The convolutional layers in YOLO extract features from the input image at multiple scales.

• These features represent different levels of abstraction, capturing both low-level and high-level visual patterns. 50 Here YOLO predicts bounding boxes for potential objects in the image.

• Each bounding box consists of coordinates (x, y) for its center, width, and height.

YOLO predicts multiple bounding boxes per grid cell, each with different sizes and aspect

ratios. For bounding boxes the values of x, y, width, height are:

x	y	width	height
826	234	69	84
903	196	83	74
1123	203	63	83
1290	288	81	68

\ Along with bounding boxes, YOLO predicts the probability of each class being present in each bounding box.

•Classes include common objects like cars, pedestrians, and in this case, an airplane. YOLO assigns a confidence score to each predicted bounding box,

• Indicating the likelihood that the box contains an object of any class. This score is a combination of the probability of the class and how well the bounding box fits around the object. The confidence score for each bounding box prediction are: Bounding box 1: Confidence Score = 0.85 Bounding box 2 : Confidence Score = 0.92 Bounding box 3 : Confidence Score = 0.78 YOLO applies non-maximum suppression to remove redundant and

• Overlapping bounding boxes. It keeps the bounding box with the highest confidence score for each detected object and eliminates others that have significant overlap with it. The performance is evaluated by mean average precision and here

• we got mean average precision as 91.5% which is very much better and give the final output accurately.

The final output of YOLO consists of bounding boxes that enclose detected objects, along with their corresponding class labels and confidence scores as shown in the figure. In this case, it include a bounding box around the airplane in the image, indicating its position of it being an airplane.

Fig 5 : Output Image

5. CONCLUSION

In this article, an improved detector ABNet with three upgrades based on Faster RCNN is proposed for RSIs. First, to explore correlations between local cross-channels, the EECA mechanism is designed to achieve more effective channel feature extraction capability for Res Net. EECA mechanism highlights the large objects in RSIs and suppresses negative information of complicated background. Second, AFPN is developed, which only introduces an MLP, a

pointwise convolution, and a nonlocal block to integrate feature maps of various scales efficiently.

Third, CEM is deployed to combine the deepest-level features of the backbone into AFPN and coalesce sufficient contextual information. Experiments on three public benchmarks prove that ABNet significantly outperforms many state-of-the-art algorithms. Our method only introduces less than 1.5M extra parameters than baseline, which maintains a decent running speed. We find that the detection performance of ABNet for small objects is not significantly improved. Therefore, how to design a lightweight and better detector for small objects will be further investigated in our future work. In addition, we will explore the performance of the EECA mechanism in other remote sensing tasks.

6.FUTURE SCOPE

Multiscale object detection in remote sensing imagery is a rapidly evolving field with significant potential for future advancements.

• Current methods struggle with small objects in high-resolution imagery. Future research will focus on techniques to better capture these objects and improve detection accuracy.

• Training deep learning models often requires vast amounts of labeled data, which can be scarce for remote sensing tasks. Future work will explore transfer learning techniques to leverage knowledge from existing models and reduce data dependency.

• Integrating data from various sensors (optical, LiDAR etc.) can provide richer information. Future research will focus on methods to effectively combine information from these sources for improved object detection.

• While deep learning models achieve high accuracy, their inner workings are often opaque. Future research will aim to develop models that are more interpretable, allowing us to understand how they arrive at their detections.

• Many applications require real-time analysis of remote sensing data (e.g., disaster response). Future research will focus on optimizing algorithms for faster processing and enabling real-time object detection.

• These advancements will lead to more robust and versatile object detection methods, unlocking a wider range of applications in remote sensing, from precision agriculture and urban planning to environmental monitoring and disaster management.

7. REFERENCES

[1] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 230–241, Jan. 2017.

[2] W. Xie, J. Lei, S. Fang, Y. Li, X. Jia, and M. Li, "Dual feature extraction network for hyperspectral image analysis," Pattern Recognit., vol. 118, Apr. 2021, Art. no. 107992.

[3] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, "Hyperspectral pansharpening with deep priors," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 5, pp. 1529–1543, May 2020.

[4] G. Ganci, A. Cappello, G. Bilotta, and C. Del Negro, "How the variety of satellite remote sensing data over volcanoes can assist hazard monitoring efforts: The 2011 eruption of nabro volcano," Remote Sens. Environ., vol. 236, Jan. 2020, Art. no. 111426.

[5] W. Xie, X. Zhang, Y. Li, J. Lei, J. Li, and Q. Du, "Weakly supervised lowrank representation for hyperspectral anomaly detection," IEEE Trans. Cybern., vol. 51, no. 8, pp. 3889–3900, Aug. 2021.

[6] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 5, pp. 1457–1470, May 2018.

[7] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," ISPRS J. Photogramm. Remote Sens., vol. 98, pp. 119–132, Dec. 2014. 58

[8] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," ISPRS J. Photogramm. Remote Sens., vol. 159, pp. 296–307, Jan. 2020.

[9] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," IEEE Trans. Pattern Anal. Mach. Intell., early access, Oct. 6, 2021, doi: 10.1109/TPAMI.2021.3117983.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017