BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

REVIEW

Deep Learning Methods Used in Remote Sensing Images: A Review

Ekram M. Rewhel[1], ***Jianqiang Li***[1], ***Amal A. Hamed***[2], ***Hatem M. Keshk***[2*], ***Amira S. Mahmoud***[2], ***Sayed A. Sayed***[2], ***Ehab Samir***[2], ***Hind H. Zeyada***[2], ***Sayed A. Mohamed***[2], ***Marwa S. Moustafa***[2], ***Ayman H. Nasr***[2], ***Ashraf K. Helmy***[2]

[1] *School of Software Engineering, Beijing University of Science and Technology, Beijing, 100083, China*

[2] *Data Reception, Analysis and Receiving Station Division, National Authority for Remote Sensing and Space Science, Cairo, 1564, Egypt*

ABSTRACT

Undeniably, Deep Learning (DL) has rapidly eroded traditional machine learning in Remote Sensing (RS) and geoscience domains with applications such as scene understanding, material identification, extreme weather detection, oil spill identification, among many others. Traditional machine learning algorithms are given less and less attention in the era of big data. Recently, a substantial amount of work aimed at developing image classification approaches based on the DL model's success in computer vision. The number of relevant articles has nearly doubled every year since 2015. Advances in remote sensing technology, as well as the rapidly expanding volume of publicly available satellite imagery on a worldwide scale, have opened up the possibilities for a wide range of modern applications. However, there are some challenges related to the availability of annotated data, the complex nature of data, and model parameterization, which strongly impact performance. In this article, a comprehensive review of the literature encompassing a broad spectrum of pioneer work in remote sensing image classification is presented including network architectures (vintage Convolutional Neural Network, CNN; Fully Convolutional Networks, FCN; encoder-decoder, recurrent networks; attention models, and generative adversarial models). The characteristics, capabilities, and limitations of current DL models were examined, and potential research directions were discussed.

Keywords: Deep Learning (DL); Satellite imaging; Image classification; Segmentation and object detection

*CORRESPONDING AUTHOR:
Hatem M. Keshk, Data Reception, Analysis and Receiving Station Division, National Authority for Remote Sensing and Space Science, Cairo, 1564, Egypt; Email: hatem.magdy@narss.sci.eg

ARTICLE INFO
Received: 3 November 2022 | Revised: 9 February 2023 | Accepted: 20 February 2023 | Published Online: 4 April 2023
DOI: https://doi.org/10.30564/jees.v5i1.5232

CITATION
Rewhel, E.M., Li, J.Q., Hamed, A.A., et al., 2023. Deep Learning Methods Used in Remote Sensing Images: A Review. Journal of Environmental & Earth Sciences. 5(1): 33-64. DOI: https://doi.org/10.30564/jees.v5i1.5232

COPYRIGHT
Copyright © 2023 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License. (https://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

The swift development in remote sensing (RS) platforms and instruments have increased the accessibility of earth observation to help Earth's surface measuring feature. Satellite platforms accumulate images at frequent intervals which results in a growing exponential volume of data. In the digital era, data become not only valuable but also intelligent. Big Data (BD) term has been introduced in mid-2011 to describe a broad set of heterogeneous large volumes of data that can hardly be managed and processed using conventional approaches. Technically, the five main dimensions that characterize BD [1] are: A massive amount of data, speed of data generation and delivery, structured and unstructured data sources, veracity, and value [2]. BD and open-source RS data open the door to improving DL approaches by extracting insights from the collected RS data to help establish more robust and effective models for RS applications.

Meanwhile, several attempts to use DL in RS have been made [3]. Scientists use contemporary technologies and different RS data sources to improve context-based feature learning and exploit the potential classification for massive volumes of remote sensing imageries.

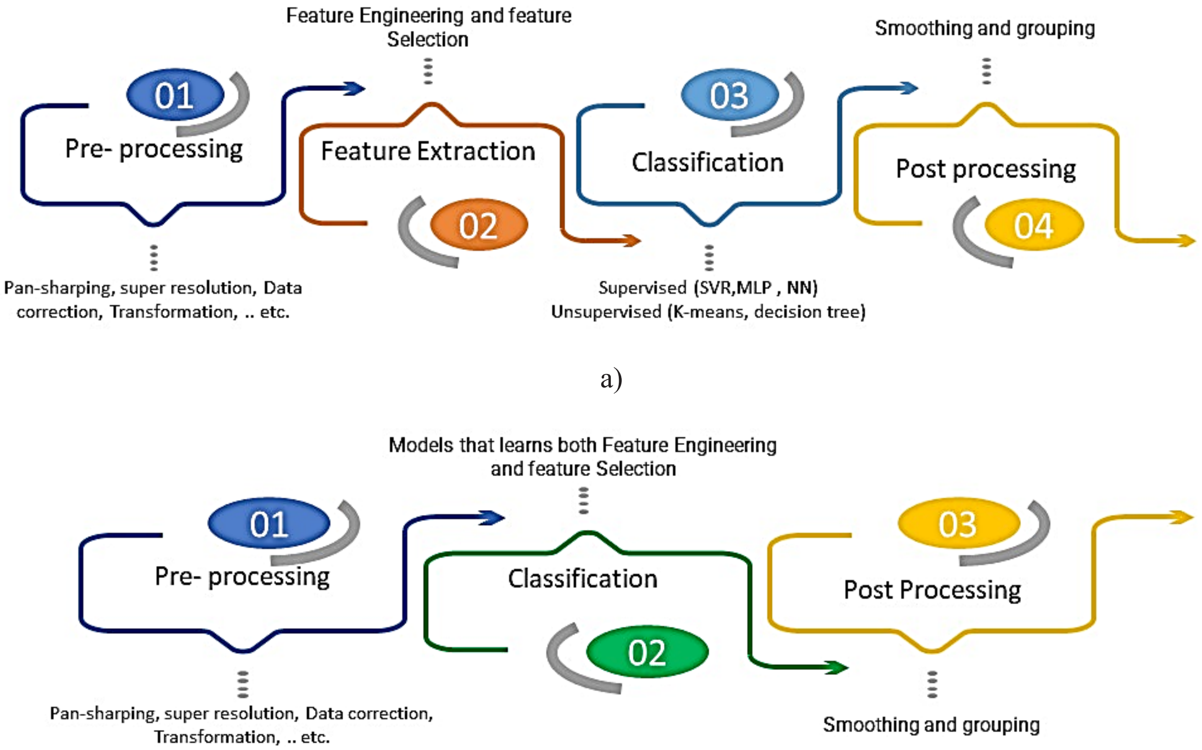
Modern RS applications [4] rely basically on image classification techniques. Typically, image classification in the remote sensing domain is grouped into supervised, non-supervised, and object-based approaches. Other criteria to group image classification are by a number of labels per image: Single and multi-label classification. RS classification pipeline [5] is composed of four main steps namely: Pre-processing, feature engineering, classification, and post-processing (see **Figure 1a**) whereas each step may include subtasks. A solid breakdown of the process into sub-tasks with specific assumptions helps develop standalone sub-problems with solutions or models that can be integrated into the classification pipeline task. The preparation process includes correcting, de-noising, and synchronizing data to increase the process performance. The feature engineering process involves removing noisy data from the input image, lowering dimensionality, and establishing a collection of suitable representations (features) for the input from which the Machine Learning (ML) model may utilize to predict the target classes. The adopted model is built based on the training samples in order to recognize the association between the training data features/representation. After training, testing, and validation, the adopted model predicts fresh data. Finally, in pixel-level classification, post-processing is a collection of procedures used to improve the final classified image [6].

Recently, the increased capability of DL has led to its use in a wide spectrum of applications in the RS domain. End-to-end architecture generalizes (**Figure 1b**) hierarchical rich feature learning. The current focus of the DL model was improved due to computing capability in new processor generations. In this context, object detection, image segmentation and scene understanding were considered typical tasks where classification approaches were empowered.

The main contributions can be summarized as follows:

- This survey analyzes the most recent publications with respect to image classification, object detection, and image segmentation problems in the remote sensing domain.
- Different DL aspects were reviewed including network architectures, loss functions, training strategies, and key contributions.
- Drawing from the latest progress by the computer vision community, several promising future directions for future research were described and how they can be integrated to value-add existing and inspire RS applications.

The rest of this article is organized as: Section 3 provides histories of remote sensing imageries. The history of deep learning architectures is summarized in Section 4. Section 5 discusses the recent efforts of deep learning in remote sensing classification, segmentation, and object detection tasks. In Section 6, the main challenges were discussed. Section 7 illustrates future directions for DL-based classification methods for Earth Observation (EO) imageries. DL-based image classification applications were highlighted in Section 8. Finally, conclusions were presented in Section 9.


Feature Engineering and feature Selection
Smoothing and grouping
01
Pre- processing
Feature Extraction
02
03
Classification
Post processing
04
Pan-sharping, super resolution, Data correction, Transformation, .. etc.
Supervised (SVR,MLP , NN)
Unsupervised (K-means, decision tree)
Models that learns both Feature Engineering and feature Selection
01
Pre- processing
Classification
02
03
Post Processing
Pan-sharping, super resolution, Data correction, Transformation, .. etc.
Smoothing and grouping

a)

b)

Figure 1. Comparison between the common steps of a) the typical machine learning approaches, and b) the modern end-to-end DL structure.

2. Overview of remote sensing imageries

Remote sensing accumulates information about an object, area, or phenomenon with no contact with it [7]. Data collection and data analysis are considered two key processes in **Figure 2**, which displays the generalized processes and elements involved in remote sensing. The data collection process includes a) energy sources, b) energy propagation through the atmosphere, c) energy interactions with earth surface features, d) energy retransmission through the atmosphere, (e) and (f) airborne and/or spaceborne sensors monitor changes in the way earth surface features reflect and emit electromagnetic energy. g) To analyze the collected data, various viewing and interpretation equipment is used. When available, the reference data (such as soil maps, crop statistics, or field-check data) are used to aid in data analysis and help in determining the extent, location, and condition acquired by the sensors. Finally, (h) and (i) the data are compiled, usually as maps, tables, or digital spatial data. Finally, the obtained information is delivered to users who utilize it to make decisions.

The latest generation of sensors produces explosion volumes of different resolution images for Earth, which created a new processing challenge. The development of an efficient image classification method for massive remote sensing imagery is critical for modern applications.

Earth observation technology is not limited to traditional platforms but extended to Light Detection and Ranging (LiDAR), and Unmanned Aerial Vehicle (UAV). As shown in **Figure 3**, the sensors are categorized as active and passive. The sun provides a convenient source of energy for remote sensing. The sun's energy is reflected, as it is for visible wavelengths, or absorbed and then re-emitted. Remote sensing instruments measure the energy that is naturally available and is called passive sensors. Some examples of passive sensors include panchromatic, multi-spectral, hyperspectral imagery. Alternatively, active sensors provide their own energy source for

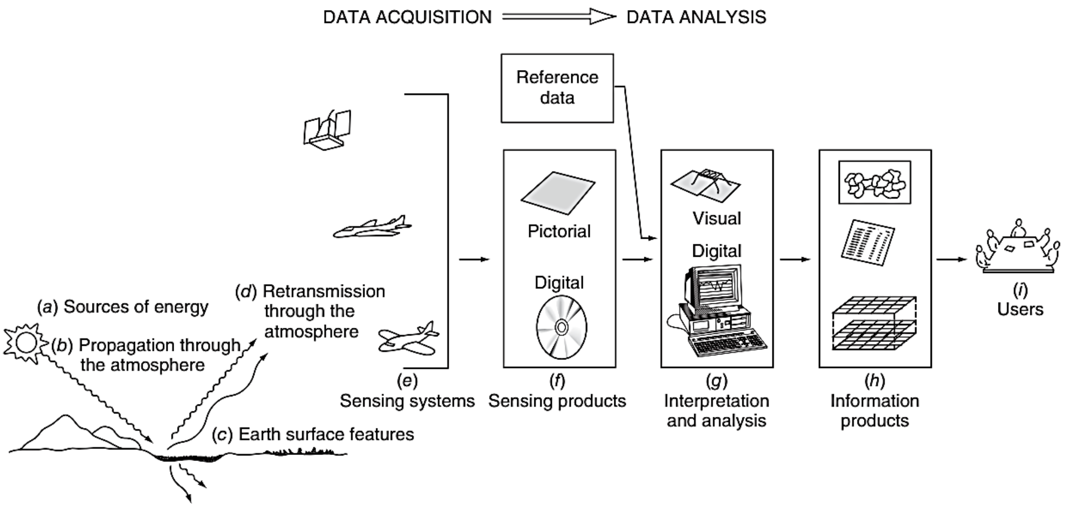
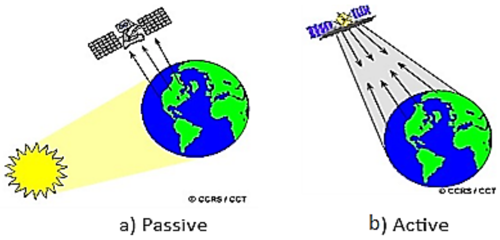

DATA ACQUISITION
DATA ANALYSIS
Reference data
Pictorial
Digital
Visual
Digital
(*a*) Sources of energy
(*b*) Propagation through the atmosphere
(*c*) Earth surface features
(*d*) Retransmission through the atmosphere
(*e*) Sensing systems
(*f*) Sensing products
(*g*) Interpretation and analysis
(*h*) Information products
(*i*) Users

Figure 2. Remote Sensing (RS) processes and elements.



a) b)

Figure 3. A graphical representation of a) passive versus b) active sensing.

illumination. The sensor emits radiation which is directed toward the target to be investigated.

The radiation reflected from that target is detected and measured by the sensor. Some examples of active sensors are LiDAR and Synthetic Aperture Radar (SAR).

Table 1 demonstrates the spatial, spectral, and temporal resolution of several common RS satellites. The spectral capabilities of the Landsat (7,8) and Sentinel (1,2) satellite missions complement each other, and with their open and cost-free access archives. The spatial resolution of images can be classified into three categories in this review: 1) High Resolution (HR) sensors (5-30 m), 2) Very High Resolution (VHR) sensors (4-m multispectral pixel size), and 3) medium to coarse resolution sensors (> 60 m multispectral pixel size). Spatial resolution is important for various applications. Coarse-resolution sensors are suitable for large-scale observation, but not for characterizing urban in compact zones. Very high- and high-resolution sensors help obtain more details.

Table 1. Specification of the common remote satellites/sensors' specifications [8].

Mission Properties	**Sentinel-2**	**Landsat 7**	**Landsat 8**	**MODIS**
Spatial resolution (m)	10, 20, 60	(15), 30, 60	(15), 30, 100	250, 500, 1000
Temporal resolution (days)	2-3	16	16	1-2
Spectral resolution	13 bands	8 bands	11 bands	25 bands
Radiometric resolution	12-bit	8-bit	16-bit	12-bit
Swath width (km)	290	185	185	2330
Wavelength range (nm)	442-2186	450-12,500	433-12,500	459-2155
Supported study area scale	local, national	national, regional	national, regional	national, regional

3. History of deep learning

This section discusses the most frequently DL architectures, including Convolutional Neural Networks (CNNs) [9], Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) [10], Encoder-Decoders (EDs) [11], and Generative Adversarial Networks (GANs) [12]. Numerous upgrades have been proposed in response to the sudden popularity growth of DL, including capsule networks, attentions, and deep belief networks. It is worth noting that in some instances, DL models are trained from scratch on new datasets (given the appropriate quality and amount of labelled data). However, transfer learning [13,14] is frequently employed to deal with incompletely labelled datasets. As illustrated in **Figure 4**, DL-based architectures were classified into eight groups based on their primary technical contributions.

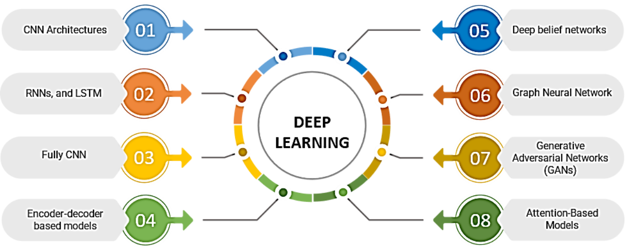


Figure 4. Deep learning architecture taxonomy.

3.1 CNN architectures

CNN family has grown since 2012, AlexNet [15] was presented at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). A standard CNN composes of three types of layers: i) convolutional layer, where filter (kernel) of weights is convolved to extract feature maps; ii) nonlinear layers, which apply an activation function on feature maps (usually elementwise) to enable the modeling of non-linear functions by the network; and iii) pooling layers, to reduce feature map spatially based on statistical information (mean, max, etc.) of a neighborhood.

Convolutional layer [16]: The input (image) is convolved then the result is passed to the next layer. Convolutional layers require four main pieces of information (filter size, number of filters, stride, and padding). The obtained result is a number of abstract feature maps equal to the number of used filters.

Pooling layer: A spatial reduction to the feature maps to minimize CNN parameters. The pooling layer has no impact on volume depth [9]. The most frequent approaches are max and average pooling [10].

Fully Connected Layer: A typical Multilayer perceptron (MLP) that transfers 2-D feature maps to 1-D vectors [11]. **Figure 5** shows that adding more parameters does not always improve precision [12]. The following sections investigate CNNs from a broader perspective.

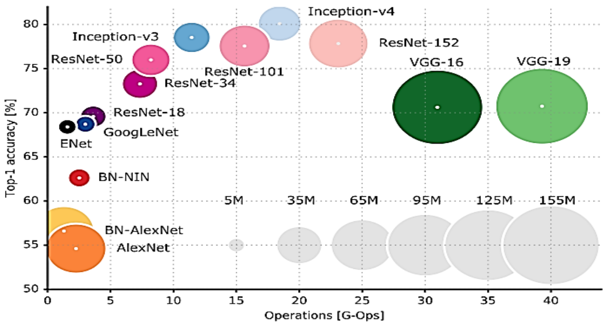
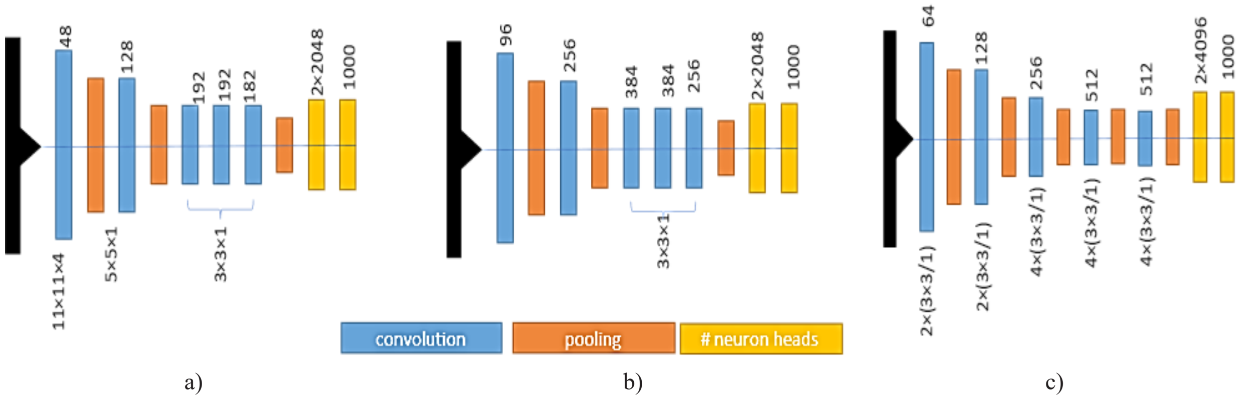


Figure 5. Summary of deep learning in terms of architecture, parameters, top-1 accuracy [12].

Vintage CNN architectures

Vintage CNN architectures include: AlexNet [15], ZFNet [17] and VGGNet [18] named after Alex Krizhevsky, Zeiler and Fergus, Visual Geometry Group [13], as shown in **Figure 6**. AlexNet is regarded as the root of CNN architectures family. The three vintage networks share a similar architecture called "template": Stacking convolution with non-linear activation followed by pooling layers to extract hierarchical features from an input image and ending with a fully connected classifier head [14]. The model provides and predicts the probability for each possible class based on the extracted features. To sum up, the main contribution of vintage architectures can be summarized as follows:

- Consisting of multiple convolutions to boost feature depth and scaling methods such as pooling with stride 2 to reduce the resolution.
- Activating the ReLU after convolutional layers speeds up backpropagation using stochas-



a) b) c)

Figure 6. Conceptual overview of the three Vintage Architectures: a) AlexNet [15], b) ZFNet [17], c) VGG-16 [18].

tic gradient descent.

- A wide range of deep networks were created by repeated building blocks such as VGG-19.

Inception family

In 2015, Google introduced a novel architecture called GoogLeNet which considered a starting point of the Inception Family and sometimes called Inception-V1. The network was built on VGG architecture in which, the Inception modules (see **Figure 7a**) with occasional max pooling to reduce the dimension (see **Figure 7b**) are stacked after the stem of the first convolutions. A typical Inception module is composed of parallel convolutions of various kernel sizes and max pooling which results in a variety of different feature maps (see **Figure 7b**). Various updated Inception versions were proposed [19-21] to boost performance using the revised sparsely connected topologies. To sum up, the Inception Family proposed a significant update to classical CNN as follows:

- Bottleneck designs and complex building block structures.
- Batch normalization to enable faster training via stochastic gradient descent.
- Factorization of convolutions in space and depth.

ResNet family

Despite their better representational ability, deeper neural networks are hard to train due to the vanishing gradient problem. As a result, the network's performance degrades dramatically as it becomes deeper. ResNet [22] was designed to facilitate training deeper neural networks and overcome the vanishing gradient problem. As shown in **Figure 8**, the primary idea of ResNet is to introduce an "identity shortcut link" that bypasses one or more layers (see **Figure 8a**). ResNet adheres to the VGG design principles while adding an identity shortcut in the residual module. Tuning a hyper-parameter is pointless because there isn't one. The pros of ResNet [23] include a) Training speed up, b) Improving the performance of classification. c) Release the power of a deeper neural network.

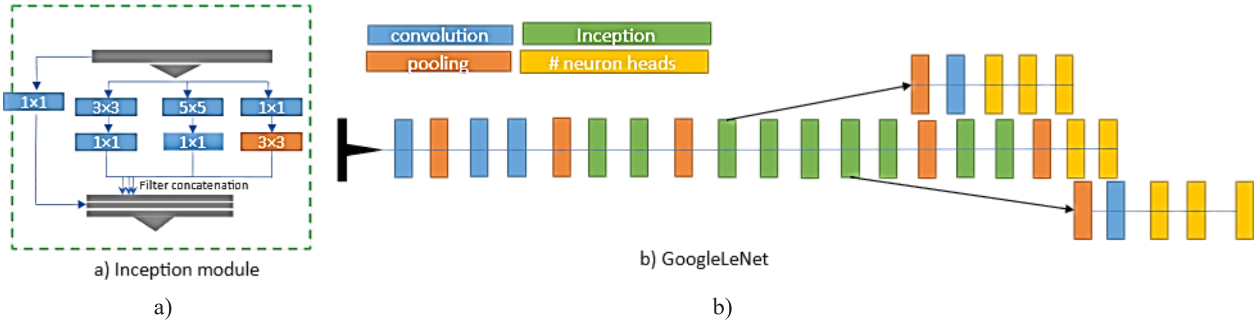


Figure 7. Conceptual overview of a) Inception module and b) Inception V-1 architecture.

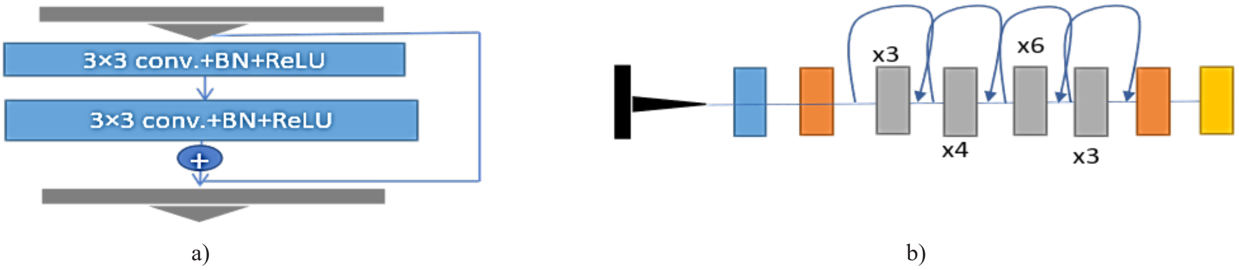


Figure 8. Conceptual overview of a) residual module and b) ResNet-50 architecture.

The architecture of ResNet has been widely investigated due to its popularity among researchers. A slew of innovative ResNet-based architectures has been revealed such as ResNeXt [23], SENet [24], SKNet [25] and ResNeSt [26].

Recent convolutions architecture

Despite their computational overhead, Vintage CNNs have shown exceptional performance in RS applications [19]. CSPNet was developed by Wang et al. [27] to reduce duplicate gradient information in the network and hence reduce inference costs. The CSPNet design reduces parameter count, increases CPU use, and reduces memory footprint [20]. CSPNet was adopted in many generic architectures such as ResNet [21], ResNeXt [24], DenseNet [23], and Scaled-YOLOv4 [26]. The CSPNet network reduces computations by 10%-20% while preserving or boosting accuracy in various recent detector types, mobile and edge devices.

Typically, the modification of the network in any of the three dimensions (depth, width, and resolution) impacted its performance. For example, increasing model depth helps capture more complex characteristics, but the model tends to become harder to train. Similarly, increasing network width captures fine-grained data but not high-level information. EfficientNet [28] is a simple architecture that uses a compound coefficient to uniformly scale all three dimensions. **Table 2** compares different deep learning models in terms of a number of parameters, accuracy.

3.2 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

Recurrent Neural Networks (RNNs) [29] are extensively employed to process temporal data (speech, text, and videos) where data at each time is dependent on prior data. While CNNs were a natural fit for 2D images, RNNs are very effective for modeling short/long-term pixel dependencies to enhance feature map estimation. Using RNNs, pixels may be linked together and processed sequentially to model global contexts and improve classification and segmentation. RNNs are unable to connect the relevant information. To handle the “long-term dependencies, Long Short-Term Memory (LSTM) [30,31] was proposed, an end-to-end Attention Recurrent Convolutional Network (ARCNet) was introduced to help focus on important high-level features in order to improve classification results.

A cascaded RNN [32] model was proposed using

Table 2. A comparison of CNN architectures.

Year	Model	Layers	Top-1 acc%	Parameters (Millions)
2012	AlexNet	7	63.3	62.4
2014	VGG-16	16	73	138.4
2014	GoogLeNet	22	-	6.7
2015	ResNet-50	50	76	25.6
2016	ResNeXt-50	50	77.8	25
2019	CSPResNeXt-50 [27]	59	78.2	20.5
2019	EfficientNet-B4	160	83	19

gated recurrent units to explore discriminative features from Hyperspectral Images (HSIs). The RNN layers help in eliminating redundant information between adjacent spectral bands and learning complementary information from nonadjacent spectral bands. An end-to-end trainable Recurrent Convolutional Neural Network (ReCNN) [33], architecture was introduced for change detection in multispectral images. The proposed architecture combined convolutional and recurrent neural networks to extract rich spectral-spatial feature representations and evaluate temporal dependency. A Siamese network based on a multi-layer Recurrent Neural Network (RNN) (SiamCRNN) [34], was designed to handle multisource multitemporal images to detect changes. SiamCRNN is an integration of three subnetworks: Deep Siamese Convolutional Neural Network (DSCNN), multiple-layers RNN (MRNN), and Fully Connected (FC) layers.

A Gated Recurrent Multi-Attention Neural Network (GRMA-Net) [35], was proposed to collect spatial informative features sequences from multi-spectral images afterward fed to a Deep-Gated Recurrent Unit (GRU) to capture long-range dependency and contextual relationship.

3.3 Fully convolutional neural network

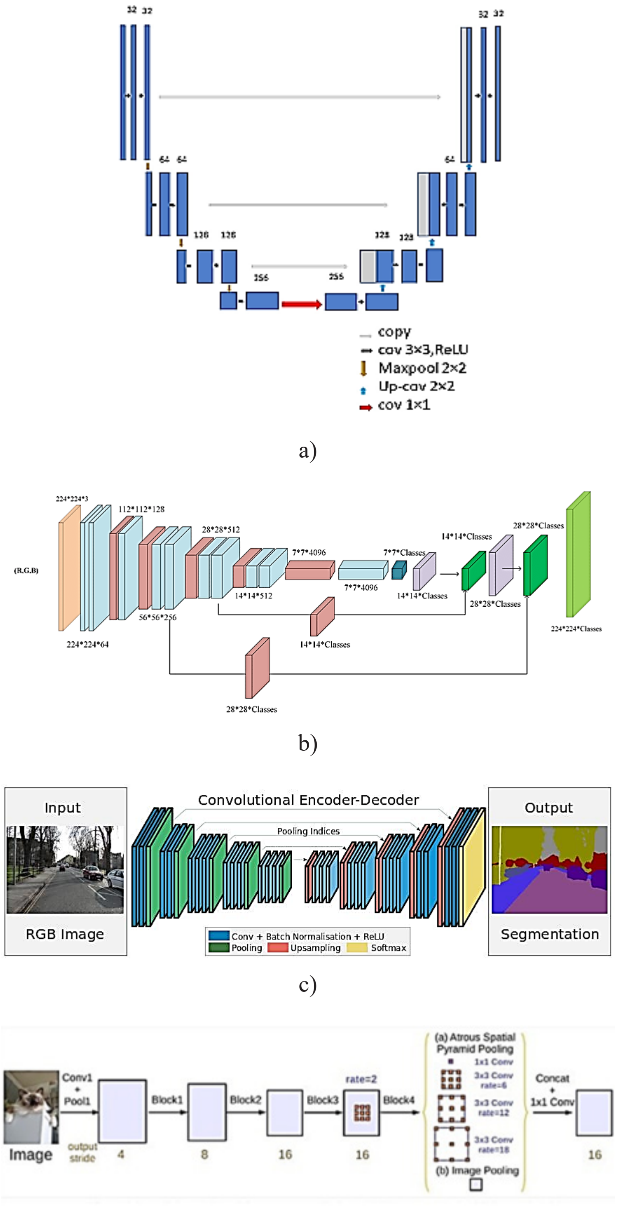
To achieve a pixel-based classification, segmentation approaches based on a Fully Convolutional Network (FCN) were proposed [36]. FCN, inspired by VGG architecture (see **Figure 9a**), contains three fundamental layers: Multi-layer convolution, deconvolution, and fusion. The fully connected layer in VGG was replaced by Convolutional layers. To compute a score for each class, a 1 × 1 convolution is adopted. The output is smaller than the input image due to pooling procedures after the convolutional layers. Deconvolution is used to bilinearly upsample these coarse outputs to regain the original image size. It works similarly to convolution but "enlarges" the input by padding the matrix and combining parts within a deconvolution filter. The deconvolution stride is inversely proportional to the upsampling factor. Deconvolution produces a scaled label segmented image. Although deconvolution recovers the original image's size, the class scores are diluted, and features are lost. To create the final segmentation, a skip architecture combines semantic information collected from a deep layer with location details from its preceding levels. The upsampled deep layer is added element-by-element to the shallow layer output.

3.4 Encoder-decoder and auto-encoder models

U-Net [37] (see **Figure 9b**) was originally designed to segment biological images. It consists of two symmetric blocks namely: the encoder and decoder. The encoder network is constructed on the basis of the FCN architecture to capture image features map. The decoder network, on the other hand, upsampled the derived feature map while lowering the number of filters. The encoder block of the original U-Net design comprises two 3 × 3 convolutions and a 2 × 2 max pooling operation with stride 2 in the encoder block. As a result, the feature map is gradually downsampled while the number of feature channels is increased. Correspondingly, the decoder block gradually raises the spatial resolution by up-sampling the feature map at each step, and then applies 2 × 2 convolution ("up-convolution") to lower the number of feature channels. To further reduce information loss, at each step of the decoder, the up-sampled feature map is concatenated with high-resolution features from the corresponding step of the encoder to avoid information loss. This is followed by two consecutive 3 × 3 convolutions that halve the feature map channel dimension. Finally, a 1 × 1 convolution is employed in the decoder's output to map the feature vector of each pixel to the appropriate number of classes, producing a pixel-wise mask.

SegNet [38] (see **Figure 9c**) incorporates two sub-networks: encoder and decoder. The encoder network uses convolution and max pooling to extract features, similar to FCNs. This network's deeper layer extracts semantic meanings. SegNet maintains the element index (i.e., the location of an element within the filter window) and uses it in the decoder network's upsampling process. Like the encoder network, the decoder network is symmetric. It translates low-resolution features to higher-resolution ones using convolutions and guided upsampling with the encoder network's pooling index. For example, a 2 × 2 low-resolution feature map becomes a 4 × 4 zero-filled matrix. This reuse of the pooling index improves boundary precision and recovers spatial information. Unlike U-Net, SegNet does not feed extracted features to decoders, which are then concatenated into upsampled feature maps.

DeepLab [39] (see **Figure 9d**) applies "Atrous convolution" with upsampled filters for dense feature extraction. Furthermore, Atrous spatial pyramid pooling encodes objects and visual context at many scales. The authors used deep convolutional neural networks and fully connected conditional random fields to yield semantically correct predictions and comprehensive object segmentation maps.



a)

b)

c)

d)

Figure 9. Fully CNN architecture a) FCN-8, b) UNet, c) SgNet, and d) DeepLab.

3.5 Deep belief network

The Deep Belief Network (DBN) [40], shown in **Figure 10**, is a subtype of Deep Neural Network made up of stacked layers of Restricted Boltzmann Machines (RBMs). It is a generative model that Geoffrey Hinton introduced in 2006 [41]. DBN may be used to solve unsupervised learning problems in order to reduce the dimension of features, as well as supervised learning tasks in order to construct classification or regression models. Two phases are required to train a DBN: Layer-by-layer training and fine-tuning. The terms "layer-by-layer training" relate to the unsupervised training of each RBM, while "fine-tuning" refers to the employment of error back-propagation techniques to fine-tune the parameters of the DBN following the unsupervised training.

Hinton suggested stacking RBMs on top of each other to train DBN quickly. During training, the lowest level RBM learns the data distribution. By sampling the previous hidden layer's hidden units, the following RBM block learns high-order correlation between them. This is done for each concealed layer up to the top.

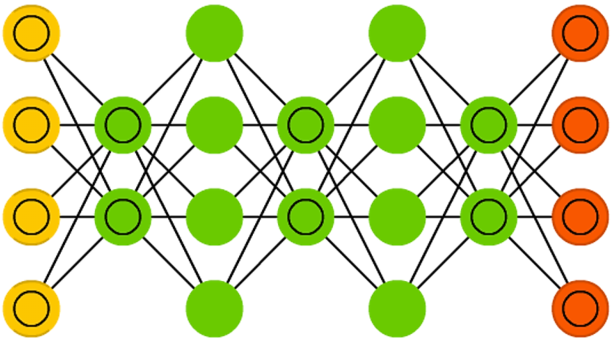


Figure 10. Graphical abstract of deep belief neural network.

3.6 Graph Neural Network (GNNs)

Graph, a data structure, represents a set of objects (nodes) and their connections (edges). Recent studies on graphs with machine learning have gained popularity due to their ability to represent a wide range of systems in different fields such as social science, natural science (physical systems) and protein-protein interaction networks [42]. Graph analysis is adopted for non-Euclidean data format in machine learning. Typical CNNs operate only with standard Euclidean data like images (2-D grids) and text (1-D sequences). Therefore, Geometric DL is the extension of deep neural models to the non-Euclidean setting. Recently, Graph Neural Network (GNNs) has recently gained popularity due to their superior performance.

Conventional CNNs are inefficient at handling spatial vectors. However, it can only be modeled as graph structures. First graph Fourier transform and convolution theorem [43], were adopted to convert vector data from the vertex domain into a point-wise product in the Fourier domain. Then, a Graph Convolutional Neural Network (GCNN) model was introduced for building pattern classification. The obtained results confirmed a satisfactory result in identifying regular and irregular building patterns. A further improvement could be considered in the potential analysis of graph-structured spatial vector data. In their pioneer work [44], a novel two-stream architecture combining global visual and object-based location features is established to enhance feature representation capabilities. First, CNN was used to extract visual features from a scene image. To learn spatial position attributes of ground objects based on GCN. The proposed architecture examines object dependencies in remote sensing scene classification for hyperspectral data.

An attempt to tackle multi-label RS image classification. This research provides a revolutionary DL-based framework called MLRSSC-CNN-GNN [45]. Basically, CNN is used to learn visual perception and create high-level appearance attributes. Each scene graph is built using the trained CNN, with nodes representing super-pixel portions of the scene. The multi-layer-integration Graph Attention Network (GAT) model is proposed to handle Multi-Label Remote Sensing Image Scene Classification (MLRSSC), where the GAT is one of the latest advancements in GNN. Extensive trials on two public MLRSSC datasets show that the proposed approach outperforms other approaches.

Several Graph Convolutional Networks (GCNs) [46], were investigated to analyze RS images to better understand their semantics which could be effective in land cover mapping. The simplification of the complexity, and the optimal control of the number of influential neighbors of the nodes were serious challenges.

High-order graph convolutional network was adopted for remote sensing scene categorization (H-GCN) [47]. During CNN feature learning, the proposed method incorporates an attention mechanism to focus on critical image components. An advanced graph convolutional network is used to analyze class dependencies (see **Figure 11**). An attentive CNN feature from each semantic class describes each node in the graph. It is possible to obtain a more informative representation of nodes by blending neighbour information of nodes in different orders. The discriminative feature representation for scene classification eventually combines H-GCN and attention CNN node representations. A summary of the current application of GNN in the RS domain is illustrated in **Table 3**.

Table 3. A summary of GNN for image analysis applications.

Summary	**Architecture**	**Application**
In object detection, and region classification. GNNs are used to calculate interested features, and region classification respectively.	Graph Attention Network	Object detection
	Graph Neural Network	Object detection
	Graph CNN	Classification
In Semantic segmentation, GNN is utilized to handle regions in images which are often not grid-like and need non-local information	Graph LSTM/Gated Graph Neural Network / Graph CNN/Graph Neural Network	Semantic Segmentation

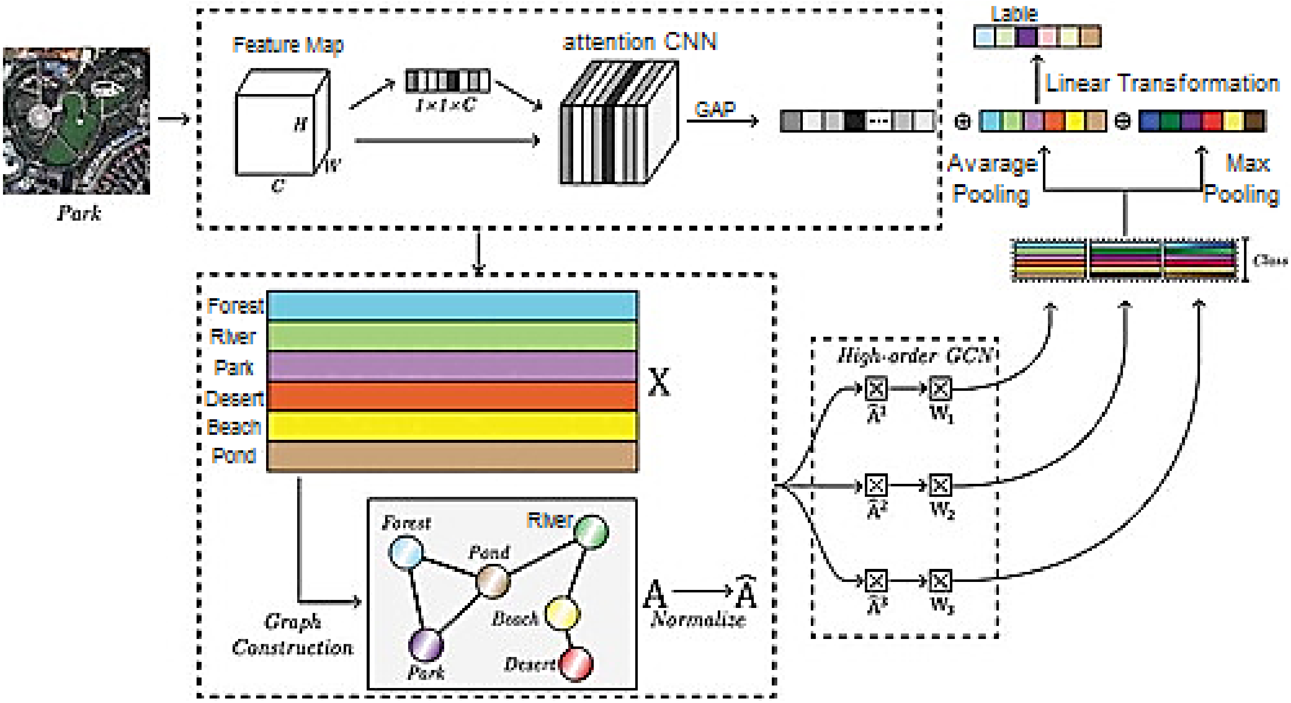


Figure 11. Scene classification framework of [47] method.

3.7 Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) [48], were presented as a novel technique for general data samples that simulate the original data distribution. Typically, GAN network is comprised of two sub-networks: Generative (G) and Discriminative (D). The Generative Network (G) maps a nonlinear function between random vectors and the desired image space. Under other conditions, the Discriminative Network (D) differentiates whether the produced data belong to the probability distribution of real data. Theoretically, GANs are built upon a game theoretical scenario whereas the generator had to compete against a discriminator [40].

Figure 12 depicts a GAN general structure. The generator network (G) directly generates samples shares training data distribution using random noise (v). The discriminator network (D) seeks to distinguish between samples from the training data and those from the generator. While the discriminator D is taught to maximize $\log\left(D\left(G(x)\right)\right)\log\left(D\left(G(x)\right)\right)$. The generator block G is trained to minimize $\log\left(1-D\left(G(z)\right)\right)\log\left(1-D\left(G(z)\right)\right)$[42]. D and G thus play a two-player minimax game as:

$$GAN = arg\min_{D}\max_{G} L_{GAN}\,(G,D)$$

where $L_{GAN}(G,D) = E_{x\sim p(x)}[\log D(x)] + E_{z\sim(z)}[\log\left(1-D\left(G(z)\right)\right)]$

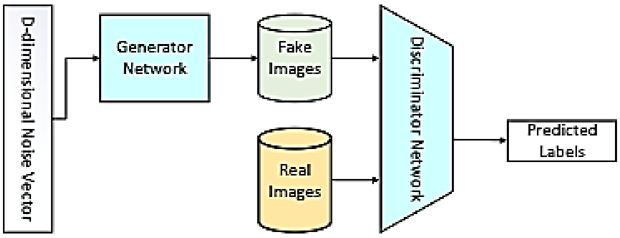


Figure 12. A typical GAN architecture in classification context.

To tackle hyperspectral challenges, the authors presented the Caps-TripleGAN framework [49], that integrated generative adversarial and CapsNet. The proposed end-to-end framework utilized a 1-D structure for sample generation. Another work introduced adversarial learning and the Variational Autoencoder (VAE) [50] was integrated to effectively classify hyperspectral imagery. The proposed method employed a conditional variational autoencoder with an adversarial training process to produce a spectral sample.

Since the introduction of GANs different types had been proposed such as conditional GAN. The authors offer new sample weighting and class adversarial training algorithms that combine SAR complementary characteristics [51]. A distribution and structure match auxiliary classifier generative adversarial network (DSM-ACGAN) was built. In DSM-ACGAN class adversarial training, statistical distribution and spatial structure are concurrently explored. DSM-ACGAN, on the other hand, uses real SAR image features to train generative models for each category. However, it also instructs the discriminator to capture both statistical and structural aspects. Class adversarial processing increases discriminative feature learning and adds to classification. It is also possible to generate class-balanced samples.

An improved GAN framework incorporated with Autoencoder (AE) to extract features advances semi-supervised and unsupervised learning [52]. The extracted features are combined with Multi-Classifier (MC) for better context classification. SAR multi frequency bands (L, C and X-bands) were effectively classified demonstrating the superiority of the proposed framework in terms of training stability and mode preservation. The authors presented GAN for land cover classification for different sources. The proposed GAN utilized FCN network for pixel classification of land covers [53]. A sea fog detection approach using Super-Pixel-Based Fully Convolutional Network (SFCNet) [54], named SFCNet was introduced. A fully connected Conditional Random Field (CRF) model was integrated to map the pixels' dependencies.

3.8 Attention-based models

Recently, attention has become a key term in DL architectures thanks to its ability to simulate human biological systems by focusing on distinct sections when processing enormous volumes of data. This section provides an overview of recent attention models. So far, DL has been difficult to interpret due to the lack of interpretability in practical and ethical concerns. The attention mechanism [55,56] helps to give distinct information with varying weights. Giving larger weights to important data draws the DL model's attention to it. Typically, existing attention models can be categorized based on four criteria: Softness of attention, input feature types, input representation, and output representation (see **Figure 13**).

Accordingly, attentions are grouped in the RS domain [57] into two main types namely: Channel and spatial, as shown in **Figure 14**. A new deep learning framework, named aTtentive weAkly Supervised Satellite image time sEries cLassifier (TASSEL) [58], was introduced to tackle time series land cover mapping. The proposed framework utilizes multifarious information instead of aggregating item statistics via the integration of graph attention mechanism and self-attention mechanism. A Spectral-Spatial Self-Attention Network (SSSAN) for HSI classification was proposed [59]. The proposed architecture is composed of two subnetworks namely spatial and spectral. The spatial self-attention module is integrated into the spatial subnetwork to enrich patch-based contextual information about the center pixel. On the other hand, a spectral self-attention module was integrated into the spectral subnetwork to take use of long-range spectral correlation over local spectral features.

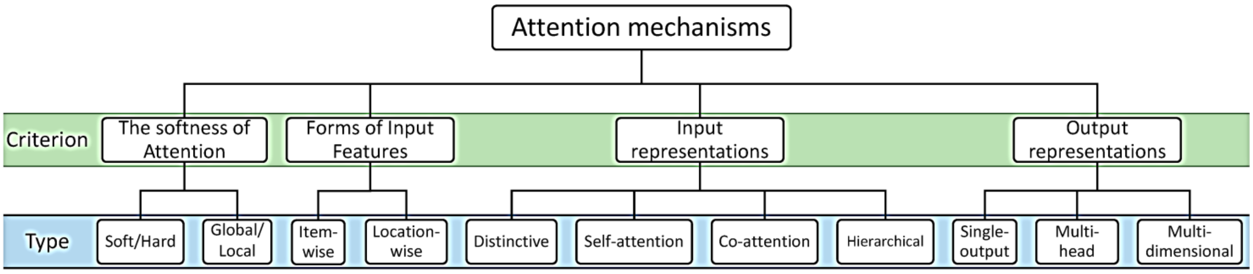


Figure 13. Several typical approaches of attention mechanisms [56].

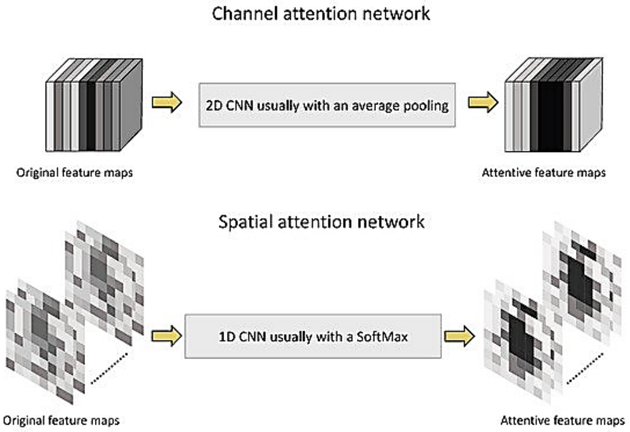

Channel attention network
2D CNN usually with an average pooling
Original feature maps
Attentive feature maps
Spatial attention network
1D CNN usually with a SoftMax
Original feature maps
Attentive feature maps

Figure 14. A simple illustration of the channel and spatial attention types/networks, and their effects on the feature maps [57].

3.9 Deep learning optimization techniques

Traditional machine learning has traditionally avoided the general optimization complexity by carefully crafting the objective function and constraints to ensure the convexity of the problem of optimization. In training neural networks, the general no-convex situation had to be addressed. This section outlines the most influential challenges involved in optimizing deep model learning such as:

Local Minima: The grandfather of all optimization problems. The local minima is a permanent challenge in the optimization of any deep learning algorithm. The local minima problem arises when the gradient encounters many local minimums that are different and not correlated to a global minimum for the cost function.

Inexact Gradients: Many deep learning models in which the cost function is intractable force an inexact estimation of the gradient. In these cases, the inexact gradients introduce a second layer of uncertainty in the model.

Flat Regions: In deep learning optimization models, flat regions are common areas that represent both a local minimum for a sub-region and a local maximum for another. That duality often causes the gradient to get stuck.

Local vs. Global Structures: Another very common challenge in the optimization of deep learning models is that local regions of the cost function don't correspond with its global structure producing a misleading gradient.

The most popular optimization method for deep learning is Stochastic Gradient Descent (SGD) [60]. The gradient estimates downwards. The learning rate is an important element in SGD. The learning rate must be decreased gradually in practice. The learning rate is one of the most difficult hyperparameters to establish in neural networks since it affects model performance. It uses a heuristic method to modify individual model parameter learning rates during training [61]. The concept is simple: If the partial derivative of the loss is positive, the learning rate should be positive. This should slow learning if the partial derivative changes sign. Examples include Adaptive Gradient Algorithm (AdaGrad), Root Mean Square Propagation (RMSProp), and Adaptive Moment Estimation (Adam).

The AdaGrad algorithm individually adapts all model parameters to their learning rates by inversely proportionally scaling them to the square root of the sum of all the squared historical gradient values [62]. The parameters with the largest partial derivatives of the loss decreased their learning rate rapidly while the parameters with small partial derivatives decreased their learning rate slowly. The RMSProp algorithm [63] modifies AdaGrad to improve non-convex performance by changing the accumulation of gradients to an exponentially weighted moving average. In a convex function, AdaGrad is designed to converge rapidly. Empirically, RMSProp has been shown to be an efficient and practical optimization algorithm for deep neural networks. Another adaptive algorithm for optimizing the learning rate is Adam [64].

4. Deep learning in remote sensing

As mentioned before, RS image classification is not limited to classification approaches, but extended to image segmentation, and object detection. This section discusses the recent efforts introduced by RS scientists.

4.1 Image classification

Recent efforts had successfully generalized to boost the performance of vintage CNNs in remote sensing classification problems. However, the insufficient number of labelled remote sensing and the complex nature of remote sensing imageries are still considered a limitation to supers the CNN performance in the remote sensing domain. Transfer learning, fine-tuning and ensemble learning were popular strategies to alleviate this limitation.

Xie et al. presented a scale-free CNN (SF-CNN) model for remote sensing scene classification [65]. The proposed architecture effectively overcomes the problem of fixed-size input images for pre-trained CNN architecture. The proposed model contains two main components: Fully Convolution Layers (FCLs) and an extra Global Average Pooling (GAP) layer. Experiments conducted on real data sets showed the superior performance of the proposed model compared with other classification methods.

In an end-to-end Feature Aggregation CNN (FACNN) was presented that utilized the intermediate features. The pre-trained VGG-16 model was adopted as a backbone to extract the intermediate features and then fed to the feature encoding module. To obtain discriminative scene representation, the classic SoftMax classifier is employed to obtain the semantic labels from the scene representations. An end-to-end learning model called Skip-Connected Covariance (SCCov) network was introduced for scene classification [66]. Skip connections and covariance pooling are embedded into the traditional CNN model. To achieve a more representative feature, skip connections architecture allows multi-resolution feature maps to combine together, and the covariance pooling to fully exploit the second-order information contained in such multi-resolution feature maps. The proposed architecture has only 10% of the parameters used by its counterparts. Experimental results demonstrate the effectiveness of the proposed model compared with the state-of-the-art techniques.

Fang et al. [67] introduced a feature representation method that incorporates frequency domain with traditional space domain. A weight spatial pyramid matching scheme was investigated to improve the performance of classification [68]. Several experiments on benchmark datasets demonstrate the superior performance of the proposed algorithm. Liu et al. introduced Siamese CNN, which combined the identification and verification models of CNNs. In addition to a metric learning regularization term imposed through CNNs to enforce more robust with the Siamese networks [69].

A bidirectional adaptive feature fusion strategy was investigated [70]. Deep features and the SIFT features were extracted using CNN and SIFT filters respectively, then fused both features to obtain a more discriminative representation and overcome the scale and rotation variability with the usage of the SIFT feature. Zhang et al. [71] proposed a new architecture named CNN-Caps Net. The proposed architecture has two parts. The first part is a pre-trained VGG-16 whose intermediate convolutional layer is utilized as a primary feature extractor. In the second part, the extracted features are fed into CapsNet. To overcome the scarcity of labelled samples, unsupervised learning-based generative adversarial networks [72] were introduced to generate training samples instead of augmentation techniques.

4.2 Image segmentation

Various efforts had been conducted to integrate the recent DL semantic segmentation techniques in the RS domain. DL image segmentation models in computer vision have been on the rise since 2014, as seen in **Figure 15**. An adaptive mask Region-based Convolutional Network (Mask-RCNN) [73] is developed for multi-class object detection in remote sensing images. Data augmentation, and transfer learning were used to address a variety of scales, sizes, and densities of remote-sensing objects. Another effort was developed [74], to extract crops from satellite imageries based on Mask RCNN. A road segmentation approach based on DeepLab v3 [75] was proposed by incorporating Squeeze-and-Excitation (SE) module in order to apply weights to different feature channels and performs multi-scale upsampling to preserve and fuse shallow and deep information. Unbalanced road samples problem in RS images, different loss functions and backbone network modules were evaluated during training.

Acone karst landscape identification based on DeepLab V3+ network [76] was proposed for multi-source data. Optical images and DEM data were used to generate the training samples. The input module of DeepLab V3+ network was altered in order to handle a four-channel image.

4.3 Object detection

Extensive studies have been devoted to studying object detection in optical and SAR images [79]. **Figure 16** illustrated the history of DL image object detection models in computer vision since 2014. Many researchers in the RS domain are using the R-CNN pipeline to recognize various geographical items in remote sensing imageries due to its superior performance in detecting natural scene image objects [77-79].

The authors [66,80] integrated a rotation-invariant CNN within the R-CNN framework for effective multi-class geospatial object detection. To further boost state-of-the-art of object detection. A novel strategy to train the CNN model called (RIFD-CNN) [81], was proposed by applying a rotation-invariant regularizer and a fisher discrimination regularizer. To accomplish precise localization of geospatial objects in HR images. Long et al. proposed an RCNN-based Unsupervised Score-Based Bounding Box Regression (USB-BBR) technique [78]. Despite the fact that the aforementioned strategies have shown to be effective in the RS community, they are nonetheless time-consuming since these methods rely on human-designed object proposal-generating methods, which consume the majority of running time. Furthermore, the quality of region suggestions developed based on hand-engineered low-level characteristics is poor, resulting in poor object identification performance.

Several studies extended the architecture of Faster R-CNN to the earth observation community [82-88]. For instance, Li et al. [84] developed a rotation-insensitive Region Proposal Network (RPN) by inserting multi-angle anchors into the existing RPN based on the Faster RCNN pipeline.

A double-channel feature combination network is also meant to learn local and contextual properties to address appearance uncertainty. Zhong et al. [85] used PSB to improve the quality of generated region proposals. For object detection, the suggested PSB framework featured FCN [36] based on the residual network [22]. The authors proposed a deformable CNN to model object changes in which non-maximum suppression [88,89] bound was established by as-

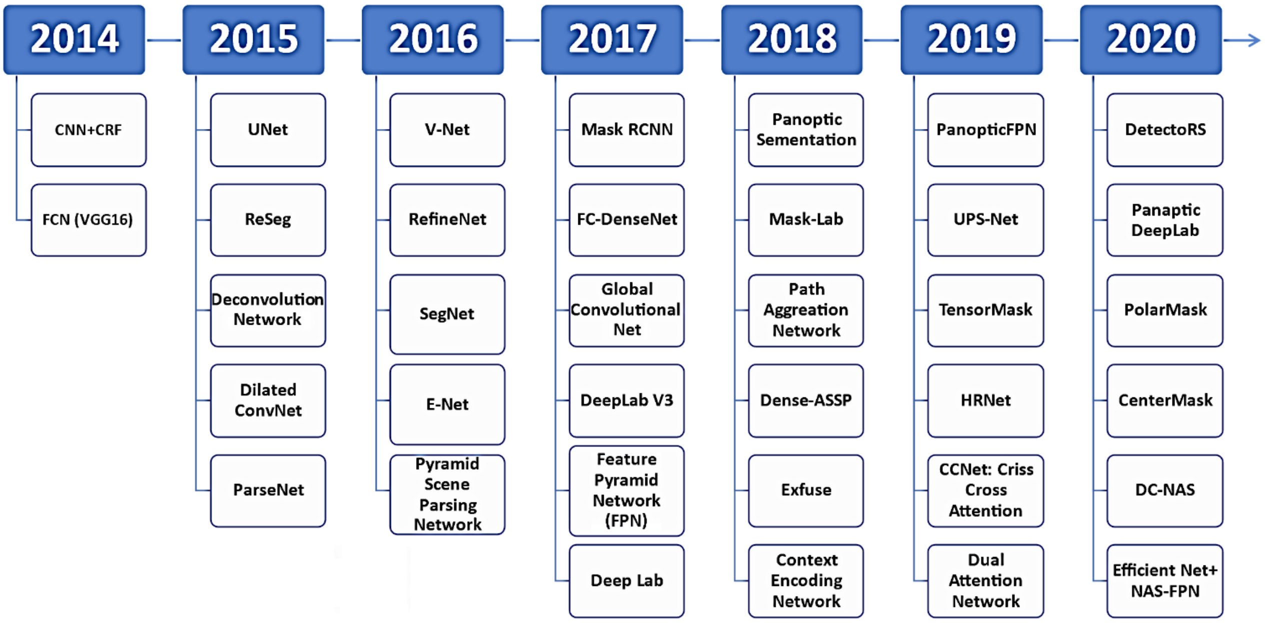


Figure 15. Timeline of representative DL-based image segmentation algorithms.

pect ratio to eliminate misleading region proposals.

To increase vehicle detection accuracy, the authors presented a Hyper Region Proposal Network (HRPN) to locate vehicle-like regions [90]. Although applying region proposal-based methods such as R-CNN, Faster R-CNN, and its variations to recognize geographical objects in Earth observation photos shows tremendous promise, amazing efforts have been made to explore other deep learning-based methods [91-95]. To determine object centroids, a rotation-invariant method [95] was employed based on super-pixel segmentation to build local patches, deep Boltzmann machines to construct high-level feature representations of local patches, and finally a series of multi-scale Hough forests to cast rotation-invariant votes. To detect ships, Zou and Shi [96] employed a singular value decompensation network to create ship-like regions, followed by feature pooling and a linear support vector machine classifier. While this detection approach is intriguing, the training method is cumbersome and slow.

Recently, some studies have attempted to translate regression-based object detection approaches developed for natural scene images to remote sensing images. Tang et al. [94] used a regression-based object detector to detect vehicle targets. Specifically, the detection bounding boxes are generated by adopting a set of default boxes with different scales per feature map location. Moreover, for each default box the offsets are predicted to better fit the object shape. Liu et al. [92] adopted a single-shot multi-box detector (SSD) framework but replaced the traditional bounding box with a rotatable bounding box from [97], in order to help to estimate objects despite their orientation angles. Liu et al. [93] developed an effective approach to detect arbitrary-oriented ships based on YOLOv2 architecture.

In addition, hard example mining [90,94], transfer learning [83], multi-feature fusion [98], and non-maximum suppression [89] are widely designed for geospatial object detection and enhance the performance of computer vision deep learning-based approaches [82]. The current stream of deep learning-based methods (e.g., R-CNN, Faster R-CNN, SSD, etc.) has proven substantial achievement in detecting geospatial objects. Earth observation photographs vary considerably from natural scene images, particularly in terms of rotation, scale variation, and complex and cluttered backgrounds [87].

4.4 Training strategies

A deep Convolutional Neural Network (CNN) can be challenging to train from scratch since it requires a significant quantity of labelled training data and much skill to guarantee that the network converges properly. Typically, feature extraction and fine-tuning of an already pre-trained network are potential options to be considered in RS (**Figure 17**).

Feature Extraction: The pre-trained CNN is employed as a feature generator. Specifically, an input image is fed to the pre-trained CNN, which then ex-

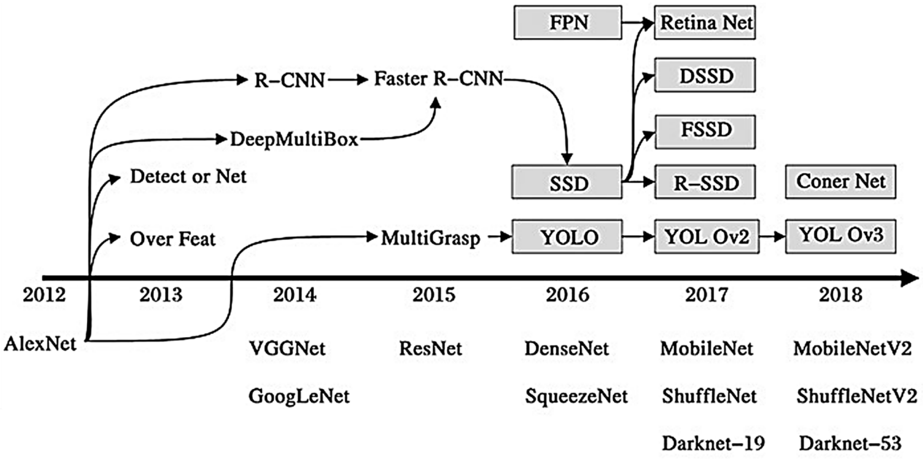
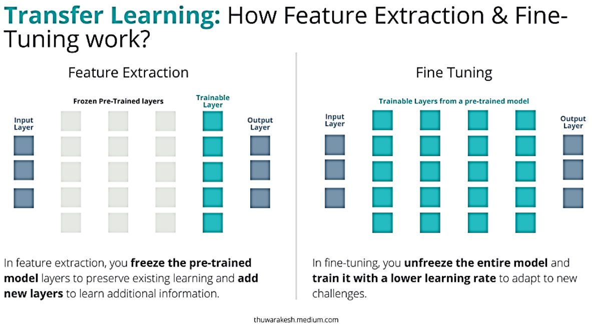


Figure 16. Timeline of progress of deep learning object detection methods.

tracts features from a specific layer of the network. The features are utilized to train a new pattern classifier. In another word, to transfer knowledge from one model to another with no training involved, the feature extraction technique is considered the key to learning features from a pre-trained model and training another (much smaller model) in order to achieve an outstanding result in a short amount of time.

Fine-Tuning: The weights of the early convolution layers are freezing while fully connected layers may be replaced with a new logistic layer relative to the application in hand. A labeled dataset is adopted to train the model while lowering the learning rate.


Transfer Learning: How Feature Extraction & Fine-Tuning work?

In feature extraction, you **freeze the pre-trained model** layers to preserve existing learning and **add new layers** to learn additional information.

In fine-tuning, you **unfreeze the entire model** and **train it with a lower learning rate** to adapt to new challenges.

thuwarakesh.medium.com

Figure 17. A comparison between feature extraction and fine-tuning training strategies.

4.5 Loss functions

Typically, the loss functions applied in image classification problems are categorized into distribution-based losses (minimize dissimilarity between two distributions), and region-based losses (minimize the mismatch or maximize the overlap regions between the two images) [99,100]. A common practice is to evaluate a small subset of the available loss function to avoid the impracticability of experimenting with all available loss functions.

Several studies compared the performance of different loss functions namely: Cross-entropy loss, focal loss, Tversky loss, dice loss, and contrastive loss to evaluate their performance in RS datasets. One can conclude that contrastive loss and weighted combined loss are widely used in RS applications due to the complex distribution of objects and their imbalance nature. **Figure 18** depicted the famous distribution-based, region-based, and compound loss functions adopted in DL for the RS domain.

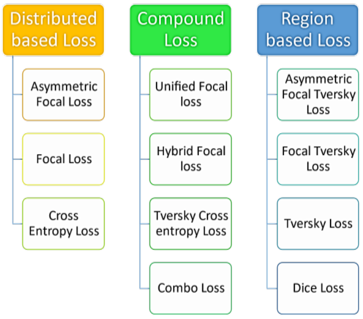


Figure 18. The famous distribution-based, region based, and compound loss functions.

4.6 Performance evaluation

Typically, the preparation of training examples is generally challenging as it requires significant labor and time to evaluate the DL performance model. Various evaluation metrics were employed that are commonly used in classification problems as described in **Table 4**. Typically, True positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN).

Table 4. Classification evaluation metrics.

Evaluation metric	Value
True Positive Rate (TPR)	$\frac{TP}{TP+FN}$
False Positive Rate (FPR)	$\frac{FP}{TP+FN}$
False Negative Rate (FNR)	$\frac{FN}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F-Measure	$\frac{2\times TP}{2\times TP+FP+FN}$
Accuracy	$100\ \times\ \frac{TP+TN}{TP+\ FN\ +\ TN+FP}$
Matthews Correlation Coefficient (MCC)	$\frac{TP\times TN\ -\ FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

5. Deep learning challenges in remote sensing domain

Undoubtedly, RS image classification has benefited tremendously from DL models. DL approaches have suppressed human-level accuracy. This section discusses the exciting challenges to tackle.

5.1 Uncertainty and balance between accuracy and efficiency

Models should establish their trustworthiness. The proposed models in the RS domain should utilize Bayesian/probabilistic inference to explicitly describe and propagate uncertainty. Identifying and treating extrapolation is also important. The contradiction between the accuracy and efficiency obtained from the models is considered a major challenge. The models with good efficiency (e.g., SegNet and ENet) fail to provide sufficiently accurate results in the RS domain.

5.2 Dependency on high-quality training data

To achieve acceptable accuracy, high-quality training datasets are required. However, collecting high-quality training data (sufficiently labeled on pixel-level annotation) is considered a hard and time-consuming task that depends on human labor.

5.3 Domain gap across different datasets

A domain gap is derived from the fact that typical deep learning models were introduced for vision tasks. The complexity of RS data impacts model performance in almost all RS applications. Since different datasets are created for different RS applications, they may differ in class number, scene look, dataset size, object size, etc. In this case, the discrepancies widen the distance between heterogeneous areas. Therefore, RS are encouraged to consider different techniques (transfer learning, data augmentation, etc.) to overcome the domain gap issue when applying DL models in their applications.

6. Recent deep learning advances in remote sensing domain

This section introduces several promising research directions to advance RS image classification algorithms.

6.1 Reinforcement learning

Reinforcement learning (RL), is an area of Machine Learning, which involves taking suitable action to maximize reward in a given scenario. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it, so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience. Combining HR images with machine learning [101], enables the scientist to address the poverty mapping problem. However, HR images come with costs and limits of scalability. RL may be utilized in combination with free low-resolution photography to dynamically identify where to gather expensive HR imagery, before doing deep learning on the HR images. Another work was introduced to utilize reinforcement learning in searching optimized parameters of deep learning model [102].

6.2 Knowledge distillation

Deep neural network shows a staining performance at the research level. However, its deployment is troublesome to utilize in limited hardware or in real-world environments due to the high computational cost and the required massive volumes of labeled data in training. To address the above problems, several model compression methods were studied to transfer the knowledge from complex architecture neural networks to compact lightweight models while sustaining performance [103].

Neural network compression is categorized into four main groups: Pruning and quantization [104], low-rank factorization [105], compact convolution filters [106], and Knowledge Distillation (KD) (Hinton et al., 2015). Pruning demands a massive number of iterations to converge to eliminate the nonessential parameters of the performance. Low-rank factorization utilized the matrix decomposition to estimate relevant parameters and remove the rest using tensor decomposition. Compact convolution filters substitute extensive parameter convolution filters with lightweight blocks. Finally, KD [103] is a simple yet effective approach to transfer or distill the representative knowledge from the large neural network (teacher) into the thin compressed network (student). The main objective of the KD is minimizing the divergence of the probabilistic outputs of teacher and student networks. The student network is trained to capture the teacher network's significant representation. Knowledge distillation [107] has been widely adopted by different architectures [108] and learning tasks [109]. Adversarial methods also have been utilized for modeling knowledge transfer between teacher and student [108].

6.3 More challenging datasets

Several large-scale RS image datasets have been created for object-based and pixel-based classification. However, more challenging datasets for different types of RS images are still required. Datasets containing a large number of overlapping objects of varying spatial resolutions would be very valuable. This can improve training models that are better at handling common scenarios in the real-world. Large-scale 3D RS image collections are in high demand because of the increasing popularity of 3D RS image datasets. These datasets are more difficult to create and effectively annotated compared with their lower-dimensional counterparts. Most 3D datasets are typically too small, and some are synthetic, therefore larger, and more challenging 3D image datasets can be extremely valuable.

6.4 Interpretable deep models

Despite the encouraging performance of DL modes, several concerns remain. For example, how and what do deep models learn? What is a minimal neural architecture that can accurately classify datasets? While methods exist to visualize the learned convolutional kernels, a detailed analysis of their behavior/dynamics is missing. A better understanding of the conceptual and theoretical aspects of the models may lead to improved models tailored to specific classification scenarios.

6.5 Weakly-supervised and unsupervised learning

Unsupervised learning and weakly supervised (few-shots) are currently hot research topics. Collecting labelled samples for RS pixel-based classification is difficult in many application areas. Transfer learning, which adopted a trained model on a large number of labelled examples (from a public benchmark), then fine-tuned on a few samples from a target application. Self-supervised learning is gaining popularity in several areas. Self-supervised learning can collect features to train efficient classification models with significantly fewer training data. Reinforcement learning models may potentially be a future approach for RS image classification.

6.6 Real-time models

Accuracy is considered a significant factor in model performance, however many applications (autonomous driving, disaster management, and land cover mapping) require running in real or near real-time. Also, some applications may be installed in limited memory and processing setting (mobile applications), but to fit them into specific devices, such as mobile phones, the networks must be simplified. Dilated convolution models, simpler models, and knowledge distillation approaches help speed up segmentation models, but there is always room for improvement.

6.7 Zero-shot learning

Zero-Shot Learning (ZSL) [110] uses the derived intermediate semantic knowledge to detect objects that have not been observed during training, which potentially extends the ability of machine learning algorithms in problem-solving skills. ZSL transmits semantic knowledge, making it an excellent complement to supervised learning. Thus, ZSL may learn to detect novel unseen classes that have no training examples by connecting them to see classes that were previously learnt. A Generalized Zero-Shot Learning

(GZSL)-based PolSAR land cover classification system is proposed [111]. Initially, basic semantic features were gathered to define typical land cover categories in PolSAR images. In the training stage, latent embedding may be used to get the projection between mid-level polarimetric information and semantic characteristics. Semantic relevance and mid-level polarimetric characteristics form the GZSL model for PolSAR data. Finally, the test instances' labels may be anticipated for some unknown classes.

6.8 UAVS, drones, and LiDAR

UAVs, as well as drones, deliver images and videos with very high-resolution amenable to be utilized in various applications [112] such as live-stock monitoring, crop production, yield prediction, and soil mapping [113]. Many sensors can be embedded in a UAV or drone, such as weather sensors, cameras, and LiDAR sensors. The obtained sensor data can be integrated into real-time decision-making in many fields [114]. LiDAR technology can create detailed topography maps and Digital Elevation Models (DEMs) necessary for land segmentation, and crop analysis field management. LiDAR technology is highly valuable in the geospatial community, with the massive data amounts amenable to utilization in a diversity of applications. Point clouds are 3D unstructured data that present many challenges for classic CNN settings. Few studies have focused on 3D point clouds. However, the 3D point cloud is gaining popularity in many applications in 3D modelling (self-driving and building modelling). Graph-based deep models may be considered as a potential area for point-clouds classification.

7. Recent deep learning in remote sensing application

This article briefly compares different deep-learning methods in the field of RS. Typically, one can observe that CNN is the most popular DL model to study and spectral-spatial features of earth observation images in classification, and object detection. The following sections review the most frequent RS applications.

7.1 Agriculture applications

Agriculture researchers have introduced some approaches, such as Transformation Learning (TL) [26] and Low Batch Learning (FSL) [27], so that deep learning models are not dependent on datasets [115]. The TL has been successfully used to identify herbs and diseases [30]. Also, FSL was found to be useful in identifying plant diseases [31-33]. The research estimates the growth stage of wheat and barley by classifying nearby images using Convolutional Neural Networks (ConvNets), and the classification was done using three different machine learning methodologies: A 5-layer ConvNet model, a transfer learning based on a VGG19 pre-trained network, and a support vector machine with conventional feature extraction [116]. Regarding the growth classification, the ConvNet learning transfer network has a much smaller training time than the built-in ConvNet model from scratch. The objectives of the research are to develop raw image-based deep learning methods for predicting the outcome in the field, and to study the sharing of multi-time images for grain quantities produced using handcrafted features and WorldView-3 and PlanetScope images, respectively [117].

7.2 Oceanography and sea ice mapping

Ocean remote sensing has reached the five-V (volume, variety, value, velocity, and veracity) age with the continual advancement of space and sensor technology over the previous 40 years. Globally, ocean remote sensing data archives top tens of petabytes, and satellite data is gathered regularly. It's difficult to harvest meaningful information from ocean remote sensing data sets. Its advantage over traditional physical or statistical-based methods for image extraction in several industrial fields has sparked interest in ocean remote-sensing applications. Two deep-learning frameworks were examined for the classification of ocean internal-wave/eddy/oil-spill/coastal-inundation/sea-ice/green-algae, and ship/coral-reef mapping [118]. SAR images were analyzed, ice charts as labelled data, and neural networks could efficiently classify ice kinds [119]. The SAR pictures were cropped into sub-regions based on the Canadian Ice Service (CIS) image analysis ice chart's latitude and longitude coordinates, and each sub-region was handled as an independent sample. Two neural networks namely: A modified U-Net and a DenseNet were adopted on the three-class dataset with dual-pol HH and HV setup, DenseNet obtained the greatest overall accuracy of 94.02 percent and ice accuracy of 91.75 percent.

For sea-ice image classification, the architecture of the SAR & optical images deep learning network was designed by extracting features and merging heterogeneous data at the feature level [120]. For the SAR images, the enhanced Spatial Pyramid Pooling (SPP) network was used and texture information about sea ice was extracted at different scales depending on the depth. As for the optical data, multilevel feature information about sea ice such as spatial and spectral information of different types of sea ice was extracted using Path Aggregation Network (PANet), which allowed the use of low-level features due to the feature of incremental extraction by the convolutional neural network. An advanced deep learning (DL) model was introduced to classify sea ice and open water from synthetic aperture radar (SAR) images [121]. U-Net was used as a backbone model for pixel-level segmentation. A DL-based feature extraction model, ResNet-34, was used as an U-Net encoder. To increase the accuracy of classifications, the original U-Net is combined with the dual attention mechanism, so as to obtain a better representation of the features, and also to form a dual attention U-Net (DAU-Net) model. The MobileNetV3 deep learning model is used as the backbone network [122], and the input samples are multi-scales, and merge the backbone network with multiscale feature fusion methods to develop a deep learning model named Multiscale MobileNet (MSMN). The MSMN accuracy was about 95% classification using SAR sea ice images and results show that dual-polarization data achieve better classification accuracy. For comparison, other classification models were trained using the training data of this paper, and the average accuracy of MSMN was found to be higher than that obtained from the model made using Convolutional Neural Networks (CNNs) and ResNet18 models. To improve classification performance, a framework for raindrop removal was introduced [123]. Images of sea ice are categorized into ice, water, ship and sky [86], by training three deep learning semantic segmentation networks, they are VGG-16, FCN, and pyramid scene parsing network. To make the training process better, transfer learning is done in addition to data augmentation. The results showed that data augmentation operations improved the performance of the three models. Also, the raindrop removal framework improves performance, the average intersection is higher than that of the VGG-16 Union.

7.3 Disaster and environmental monitoring

There is no doubt that the era of big data and deep learning has opened new options for disaster management, thanks to the diverse capabilities it provides in visualizing, analyzing, and predicting disasters. The integration of big data and DL has completely altered the strategies followed by human societies and disaster management agencies to reduce human suffering and economic losses resulting from disasters. In our world which is now mainly dependent on information technology, the main goal of computer experts and decision-makers is to make the best of model by gathering information from different sources and formats and storing it in effective ways to be used effectively in different stages of disaster management. The availability of various big data sources such as satellite imageries, Global Positioning System (GPS) traces, mobile Call Detail Records (CDRs), social media posts, etc., in conjunction with the enhancements in data analytic techniques (e.g., data mining, machine learning, and deep learning) can facilitate geospatial information extraction, that is crucial for immediate and effective disaster response. The research [124] introduced a deep neural network approach for detecting submerged stop signs in images of flooded roads and intersections, as well as detecting Canny and probabilistic

Hough transform for estimating pole length and floodwater depth. They developed a classification model using deep neural networks that successfully identified affected areas using grounded images [125]. These areas were removed from social media platforms that were downloaded immediately after the disaster. Thus, this can facilitate the acceleration of the recovery process, by marking the areas where the disaster has a greater impact than other areas.

7.4 Archaeology applications

Geospatial data and imageries are the most active field for archaeologists utilizing deep learning. Rarely can archaeology create the vast volumes of systematically coded data required for ML [126]. As a result, the increased availability of large-scale lidar, satellite, and aerial photography is changing archaeology globally, notably the finding and mapping of ancient sites. DL algorithms can analyze geographical data to find locations in various contexts. This method can determine the contribution of different variables that predict where sites are found across landscapes. Its many sizes enable archaeologists to better manage and investigate heritage at a global level. These historic landscape ML methods can help mitigate some of the challenges of predictive modelling for cultural resource management. This covers ways to assess the ML predictions' internal coherence and to investigate the factors that influence the presence or absence of archaeological sites in a landscape. This is essential in places where archaeological sites are difficult to access [127]. Two artificial intelligence approaches are introduced [128] over two areas of interest in the image processing field. They implemented a random forest classifier in their paper using the cloud platform of the Google Earth Engine data and a Single Shot Detector neural network is developed too. The final results show that this approach can be used in the future to detect scattered pottery pieces during the pedestrian archaeological survey, even if there is a great spectral similarity between the pottery and the surface of the earth. The U-Net neural network has been made to perform semantic segmentation of the data derived from airborne laser scanning cameras for the extraction of archaeological features in the Białowieża Forest in Poland [129]. The evaluation of the U-Net segmentation model is done using a pixel-level similarity measure between the ground truth and the predicted segmentation masks. The results indicated that the U-Net deep learning model is very good at a semantic segmentation of images.

7.5 Interferometry applications

While CNNs have shown high object identification accuracy in aerial pictures, few researchers have used deep-learning techniques and CNNs to identify landslides. Yu et al. [130] utilized a CNN and an enhanced region growth algorithm (RSG-R) to detect landslides. They used the RSG-R algorithm to extract discriminant information such as the area and border of landslides and determined that their CNN approach had excellent detection accuracy for detecting landslide features. Landslide identification using GF-1 images with four spectral bands and 8 m spatial resolution for Shenzhen was assessed [131]. Their automated landslide detection technique has a 72.5% detection rate, a 10.2% false positive rate, and a 67% overall accuracy. This review indicates the potential of employing CNNs for landslide detection has not yet been completely explored. CNN was adopted to identify landslides using optical satellite images from the Rapid Eye sensor (see **Figure 19**) then the obtained results were compared to state-of-the-art ML techniques, ANN, and SVM [132].

In Wenchuan Baoxing in Sichuan Province, China, images of areas where the landslide disaster occurred are captured using low-altitude unmanned aerial vehicles (UAV) for research [133]. A landslide extraction approach based on Transfer Learning (TL) model and object-oriented image analysis (TLOEL) was introduced; the TLOEL results were compared with those of the object-oriented nearest neighbor classification (NNC). It is approved that the accuracy of the TLOEL method is better than the NNC method, which helps to detect and extract finely distributed medium and small landslides, not just large landslides.

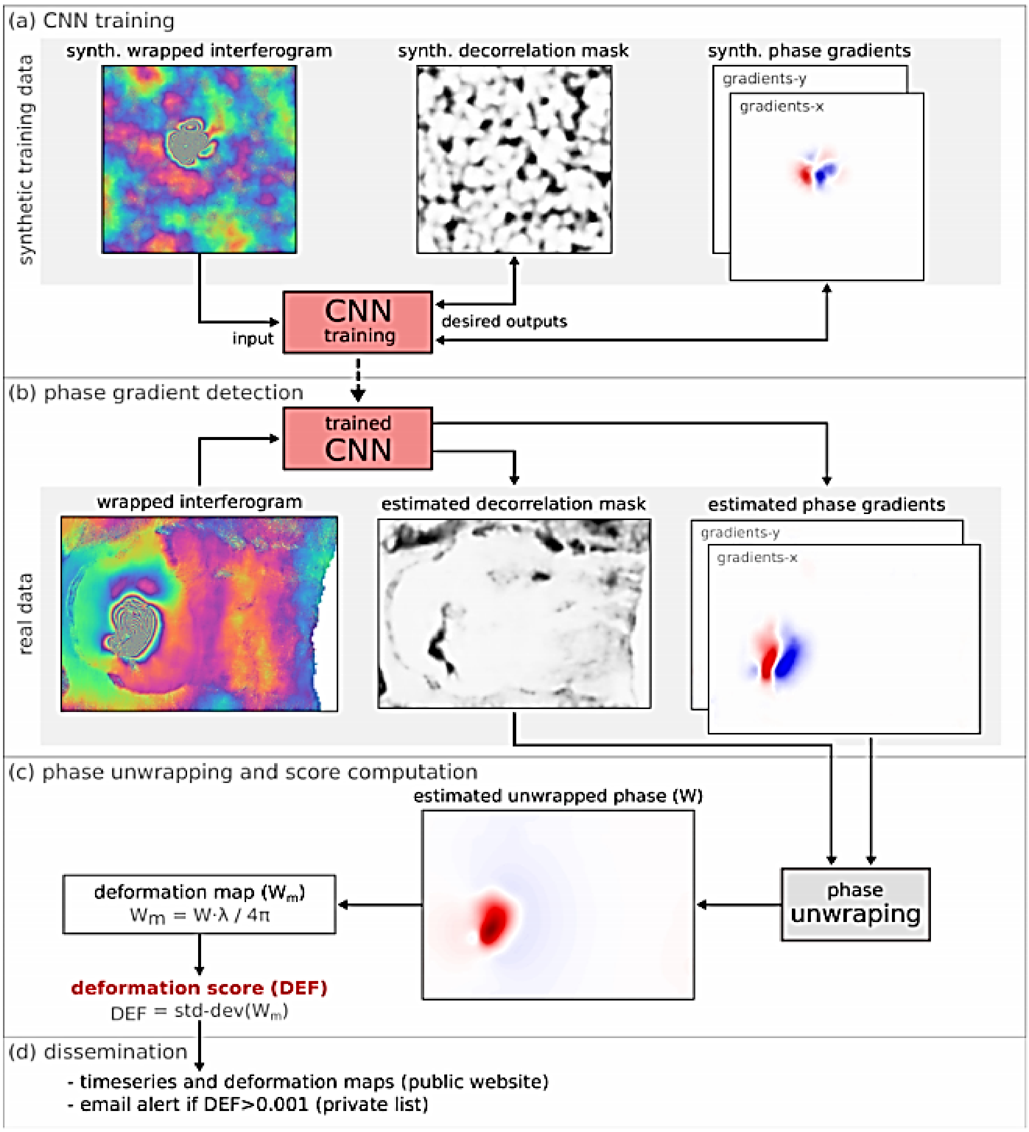

(a) CNN training
synthetic training data
synth. wrapped interferogram
synth. decorrelation mask
synth. phase gradients
gradients-y
gradients-x
input
CNN training
desired outputs
(b) phase gradient detection
trained CNN
real data
wrapped interferogram
estimated decorrelation mask
estimated phase gradients
gradients-y
gradients-x
(c) phase unwrapping and score computation
estimated unwrapped phase (W)
deformation map (W_m)
$W_m = W \cdot \lambda / 4\pi$
phase unwraping
deformation score (DEF)
DEF = std-dev(W_m)
(d) dissemination
- timeseries and deformation maps (public website)
- email alert if DEF>0.001 (private list)

Figure 19. Landslides identification using deep learning framework [132].

Another work introduced [134] for volcano deformation detection. The CNN is trained on simulated data and is later used to detect phase gradients and a decorrelation mask from input-wrapped interferograms to locate ground deformation caused by volcanoes. The paper [135] proposes the use of self-supervised contrastive learning to learn high-quality visual representations within interferometric synthetic aperture radar (InSAR) data. A SimCLR framework is achieved to find a solution based on a specialized architecture or a large classified or synthetic dataset. The self-supervised pipeline has been shown to give higher accuracy compared to the state-of-the-art methods and shows good generalization for the out-of-distribution test data also. The approach is approved for its high potential for detecting unrest episodes prior to the recent Icelandic volcanic eruption.

7.6 Climate and environmental applications

The remarkable flexibility and adaptability of deep learning models enable scientists to identify, classify and localize extreme weather events under various climate change scenarios. Several attempts had been conducted to adopt DL models to study climate and environment. ClimateNet [136] is pioneer research introduced to analyze a pixel-based detection for tropical cyclones (TCs) and atmospheric rivers (AC). Another study was conducted to develop the Optimized Ensemble Deep Learning (OEDL) framework [137] to forecast waves.

Reiersen et al. have developed a database named ReforesTree that includes data on carbon stocks in some forests in Ecuador [138]. The project aims to overcome the carbon deficiency in some interested forests. A comprehensive deep learning-based model that detects trees individually in RGB drone images has demonstrated that the forest carbon stocks can be professionally calculated according to the official standards of carbon offset certification.

Researchers [139] proposed a deep learning-based approach (i.e., U-Net) using the landscape pattern using Sentinel-1 data to produce forest harvesting maps per month within three years. The variable harvest pattern was obtained from Sentinel-1 data using the U-Net bottleneck block as the integrated entities. This modern approach is an important step in the mapping of forest harvesting at monthly intervals of forest harvesting as well as in the development of a sustainable forest management strategy to assist the beneficiaries.

The collection of remote sensing and social sensing data was studied [140] to make informational maps showing the extent of the flood. That is why deep learning methods are used to deal with heterogeneous data. Regarding remote sensing data, it turns out that the given deep learning models predict flood water much better. In the case of social sensing, two layers of data were used as related tweet text and images for the case study areas, thus heterogeneous data sources could be combined to complement each other. After analyzing the results of this study, three types of signals are defined: (1) definite signals from the two sources, which confirm that water has flooded a specific area, (2) complementary signals that give multiple information in a context such as requirements and needs, disaster outcomes or reports, and (3) New signals in the event that the two sources do not overlap and their information is not repeated. (4) Novel signals when both data sources do not overlap and provide unique information.

8. Conclusions

This article conducted a comparative review to inspect the recent cutting-edge research of DL in the remote sensing field. DL can help remote sensing scientists overcome several challenges in real-world applications, such as urban planning, natural hazards detection, environment monitoring, vegetation mapping, and geospatial object identification. However, it required a hefty investment to be integrated. This context introduced reviews in DL in RS classification, indicating DL's prominent role in tackling the RS challenges. Therefore, ample conclusions were drawn:

- The up-raising trend in adopting DL architectures in different applications, the availability of free satellite imagery, and the massive computational capabilities and efficient learning algorithms help researchers gain insights and recommend solutions to several modern challenges.
- Freely available satellite imageries were employed effectively in agriculture applications and change maps, especially Landsat and Sentinel-2 imagery.
- Extensive studies adopted different machine learning methods for RS data processing. In the last five years, DL had been adopted in several studies, especially in crop mapping and Interferometry applications.
- The use of the recent CNN advances (attention, GNN, uncertainty) for various applications has significantly increased since 2018. This increased rate in modern architectures in RS image classification highlights its effectiveness and popularity.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study was partially supported by National Authority for Remote Sensing and Space Science.

References

[1] Emani, C.K., Cullot, N., Nicolle, C., 2015. Understandable big data: A survey. Computer Science Review. 17, 70-81.

[2] Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture. 143, 23-37.

[3] Fernandez, A., Insfran, E., Abrahão, S., 2011. Usability evaluation methods for the web: A systematic mapping study. Information and Software Technology. 53(8), 789-817.

[4] Li, Y., Ma, L., Zhong, Z., et al., 2020. Deep learning for LiDAR point clouds in autonomous driving: A review. IEEE Transactions on Neural Networks and Learning Systems. 32(8), 3412-3432.
DOI: https://doi.org/10.1109/TNNLS.2020.3015992

[5] Fayyad, J., Jaradat, M.A., Gruyer, D., et al., 2020. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. Sensors. 20(15), 4220.

[6] Zhang, P., Du, P., Lin, C., et al., 2020. A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR Data. Remote Sensing. 12(22), 3764.

[7] Lillesand, T., Kiefer, R.W., Chipman, J., 2015. Remote sensing and image interpretation. John Wiley & Sons: New York.

[8] Radočaj, D., Obhođaš, J., Jurišić, M., et al., 2020. Global open data remote sensing satellite missions for land monitoring and conservation: A review. Land. 9(11), 402.

[9] Voulodimos, A., Doulamis, N., Doulamis, A., et al., 2018. Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience. 2018(Pt.I).
DOI: https://doi.org/10.1155/2018/7068349

[10] Bulo, S.R., Neuhold, G., Kontschieder, P. (editors), 2017. Loss max-pooling for semantic image segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI, USA. USA: IEEE. p. 7082-7091.

[11] Asokan, A., Anitha, J., Patrut, B., et al., 2021. Deep feature extraction and feature fusion for bi-temporal satellite image classification. CMC-Computers Materials & Continua. 66(1), 373-388.

[12] Bianco, S., Cadene, R., Celona, L., et al., 2018. Benchmark analysis of representative deep neural network architectures. IEEE Access. 6, 64270-64277.

[13] Mahmoud, A., Mohamed, S., El-Khoribi, R., et al., 2020. Object detection using adaptive mask RCNN in optical remote sensing images. International Journal of Intelligent Systems. 13, 65-76.

[14] Chen, L.C., Papandreou, G., Kokkinos, I., et al., 2017. DeepLab: Semantic image segmentation with deep convolutional Nets, Atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 40(4), 834-848.

[15] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 25, 1097-1105.

[16] Yamashita, R., Nishio, M., Do, R.K.G., et al., 2018. Convolutional neural networks: An overview and application in radiology. Insights into Imaging. 9(4), 611-629.

[17] Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. European Conference on Computer Vision. 8689, 818-833.

[18] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
DOI: https://doi.org/10.48550/arXiv.1409.1556

[19] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing

internal covariate shift. arXiv:1502.03167.
DOI: https://doi.org/10.48550/arXiv.1502.03167

[20] Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (editors), 2016. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27-30; Las Vegas, NV, USA. USA: IEEE. p. 2818-2826.

[21] Szegedy, C., Ioffe, S., Vanhoucke, V. (editors), et al., 2017. Inception-v4, inception-Resnet and the impact of residual connections on learning. Thirty-first AAAI Conference on Artificial Intelligence. 31(1).
DOI: https://doi.org/10.1609/aaai.v31i1.11231

[22] He, K., Zhang, X., Ren, S., et al. (editors), 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27-30; Las Vegas, NV, USA. USA: IEEE. p. 770-778.

[23] Xie, S., Girshick, R., Dollár, P., et al. (editors), 2017. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21-26; Honolulu, HI, USA. USA: IEEE. p. 1492-1500.

[24] Cheng, D., Meng, G., Cheng, G., et al., 2016. SeNet: Structured edge network for sea–land segmentation. IEEE Geoscience and Remote Sensing Letters. 14(2), 247-251.

[25] Wu, W., Zhang, Y., Wang, D., et al., 2020. SK-Net: Deep learning on point cloud via end-to-end discovery of spatial keypoints. Proceedings of the AAAI Conference on Artificial Intelligence. 34(4), 6422-6429.

[26] Zhang, H., Wu, Ch.R., Zhang, Zh.Y., et al., 2020. ResNeSt: Split-attention networks. arXiv:2004.08955.
DOI: https://doi.org/10.48550/arXiv.2004.08955

[27] Wang, C.Y., Liao, H.Y.M., Wu, Y.H., et al., 2020. CSPNet: A new backbone that can enhance learning capability of CNN. arXiv:1911.11929.
DOI: https://doi.org/10.48550/arXiv.1911.11929

[28] Tan, M., Le, Q. (editors), 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning, PMLR. 97, 6105-6114.

[29] Yu, Y., Si, X., Hu, C., et al., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation. 31(7), 1235-1270.

[30] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation. 9(8), 1735-1780.

[31] Wang, Q., Liu, S., Chanussot, J., et al., 2019. Scene classification with recurrent attention of VHR remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 57(2), 1155-1167.

[32] Hang, R., Liu, Q., Hong, D., et al., 2019. Cascaded recurrent neural networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing. 57(8), 5384-5394.

[33] Mou, L., Bruzzone, L., Zhu, X.X., 2019. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Transactions on Geoscience and Remote Sensing. 57(2), 924-935.

[34] Chen, H., Wu, C., Du, B., et al., 2020. Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network. IEEE Transactions on Geoscience and Remote Sensing. 58(4), 2848-2864.

[35] Li, B., Guo, Y., Yang, J., et al., 2021. Gated recurrent multiattention network for VHR remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing. 60, 1-13.

[36] Long, J., Shelhamer, E., Darrell, T. (editors), 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7-12; Boston, MA, USA. USA: IEEE. p. 3431-3440.

[37] Ronneberger, O., Fischer, P., Brox, T., 2015.

U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-assisted Intervention. 9351, 234-241.

[38] Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39(12), 2481-2495.

[39] Chen, X., Girshick, R., He, K. (editors), et al., 2019. Tensormask: A Foundation for Dense Object Segmentation [Internet]. Proceedings of the IEEE/CVF International Conference on Computer Vision. Available from: https://openaccess.thecvf.com/content_ICCV_2019/papers/Chen_TensorMask_A_Foundation_for_Dense_Object_Segmentation_ICCV_2019_paper.pdf

[40] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT Press: Cambridge.

[41] Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science. 313(5786), 504-507.

[42] Fout, A.M. (editor), 2017. Protein interface prediction using graph convolutional networks. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec. p. 6533-6542.

[43] Yan, X., Ai, T., Yang, M., et al., 2019. A graph convolutional neural network for classification of building patterns using spatial vector data. ISPRS Journal of Photogrammetry and Remote Sensing. 150, 259-273.

[44] Liang, J., Deng, Y., Zeng, D., 2020. A deep neural network combined CNN and GCN for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 13, 4325-4338.

[45] Li, Y., Chen, R., Zhang, Y., et al., 2020. Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network. Remote Sensing. 12(23), 4003.

[46] Baroud, S., Chokri, S., Belhaous, S., et al., 2021. A brief review of graph convolutional neural network based learning for classifying remote sensing images. Procedia Computer Science. 191, 349-354.

[47] Gao, Y., Shi, J., Li, J., et al., 2021. Remote sensing scene classification based on high-order graph convolutional network. European Journal of Remote Sensing. 54(sup1), 141-155.

[48] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al., 2020. Generative adversarial networks. Communications of the ACM. 63(11), 139-144.

[49] Wang, X., Tan, K., Du, Q., et al., 2019. Caps-TripleGAN: GAN-Assisted CapsNet for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing. 57(9), 7232-7245.

[50] Wang, X., Tan, K., Du, Q., et al., 2020. CVA-E: A conditional variational autoencoder with an adversarial training process for hyperspectral imagery classification. IEEE Transactions on Geoscience and Remote Sensing. 58(8), 5676-5692.

[51] Ren, Z., Hou, B., Wu, Q., et al., 2020. A distribution and structure match generative adversarial network for SAR image classification. IEEE Transactions on Geoscience and Remote Sensing. 58(6), 3864-3880.

[52] Zhang, Z., Yang, J., Du, Y., 2020. Deep convolutional generative adversarial network with autoencoder for semi supervised SAR image classification. IEEE Geoscience and Remote Sensing Letters. 19, 1-5.

[53] Ji, S., Wang, D., Luo, M., 2021. Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 59(5), 3816-3828.

[54] Huang, Y., Wu, M., Guo, J., et al., 2021. A correlation context-driven method for sea fog detection in meteorological satellite imagery. IEEE Geoscience and Remote Sensing Letters. 19, 1-5.

[55] Guidotti, R., Monreale, A., Ruggieri, S., et al., 2018. A survey of methods for explaining black

box models. ACM Computing Surveys (CSUR). 51(5), 1-42.

[56] Niu, Z., Zhong, G., Yu, H., 2021. A review on the attention mechanism of deep learning. Neurocomputing. 452, 48-62.

[57] Ghaffarian, S., Valente, J., van der Voort, M., et al., 2021. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. Remote Sensing. 13(15).

[58] Ienco, D., Gbodjo, Y.J.E., Gaetano, R., et al., 2020. Weakly supervised learning for land cover mapping of satellite image time series via attention-based CNN. IEEE Access. 8, 179547-179560.

[59] Zhang, X., Sun, G.Y., Jia, X.P., et al., 2021. Spectral-spatial self-attention networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing. 60, 1-15.

[60] Robbins, H., Monro, S., 1951. A stochastic approximation method. The Annals of Mathematical Statistics. 22(3), 400-407.

[61] Jacobs, R.A., 1988. Increased rates of convergence through learning rate adaptation. Neural Networks. 1(4), 295-307.

[62] Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 12(7), 257-269.

[63] Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning. 4(2), 26-31.

[64] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
DOI: https://doi.org/10.48550/arXiv.1412.6980

[65] Xie, J., He, N., Fang, L., et al., 2019. Scale-free convolutional neural network for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing. 57(9), 6916-6928.

[66] He, N., Fang, L., Li, S., et al., 2019. Skip-connected covariance network for remote sensing scene classification. IEEE Transactions on Neural Networks and Learning Systems. 31(5), 1461-1474.

[67] Fang, J., Yuan, Y., Lu, X., et al., 2019. Robust space-frequency joint representation for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing. 57(10), 7492-7502.

[68] Liu, B.D., Meng, J., Xie, W.Y., et al., 2019. Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification. Remote Sensing. 11(5), 518.

[69] Liu, X., Zhou, Y., Zhao, J., et al., 2019. Siamese convolutional neural networks for remote sensing scene classification. IEEE Geoscience and Remote Sensing Letters. 16(8), 1200-1204.

[70] Lu, X., Ji, W., Li, X., et al., 2019. Bidirectional adaptive feature fusion for remote sensing scene classification. Neurocomputing. 328, 135-146.

[71] Zhang, W., Tang, P., Zhao, L., 2019. Remote sensing image scene classification using CNN-CapsNet. Remote Sensing. 11(5), 494.

[72] Ma, D., Tang, P., Zhao, L., 2019. SiftingGAN: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro. IEEE Geoscience and Remote Sensing Letters. 16(7), 1046-1050.

[73] Keshk, H.M., Yin, X.C., 2020. Change detection in SAR images based on deep learning. International Journal of Aeronautical and Space Sciences. 21, 549-559.
DOI: https://doi.org/10.1007/s42405-019-00222-0

[74] Wang, S., Sun, G., Zheng, B., et al., 2021. A crop image segmentation and extraction algorithm based on mask RCNN. Entropy. 23(9), 1160.

[75] Lin, Y., Xu, D., Wang, N., et al., 2020. Road extraction from very-high-resolution remote sensing images via a nested SE-DeepLab model. Remote Sensing. 12(18), 2985.

[76] Fu, H., Fu, B., Shi, P., 2021. An improved segmentation method for automatic mapping of cone karst from remote sensing data based on DeepLab V3+ Model. Remote Sensing. 13(3), 441.

[77] Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 54(12), 7405-7415.

[78] Long, Y., Gong, Y., Xiao, Z., et al., 2017. Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing. 55(5), 2486-2498.

[79] Ševo, I., Avramović, A., 2016. Convolutional neural network based automatic object detection on aerial images. IEEE Geoscience and Remote Sensing Letters. 13(5), 740-744.

[80] Deng, Z., Sun, H., Zhou, S., et al., 2017. Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 10(8), 3652-3664.

[81] Cheng, G., Han, J., Zhou, P., et al., 2018. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. IEEE Transactions on Image Processing. 28(1), 265-278.

[82] Guo, W., Yang, W., Zhang, H., et al., 2018. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. Remote Sensing. 10(1), 131.

[83] Han, X., Zhong, Y., Zhang, L., 2017. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sensing. 9(7), 666.

[84] Li, K., Cheng, G., Bu, S., et al., 2017. Rotation-insensitive and context-augmented object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing. 56(4), 2337-2348.

[85] Zhong, Y., Han, X., Zhang, L., 2018. Multiclass geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 138, 281-294.

[86] Yao, Y., Jiang, Z., Zhang, H., et al., 2017. Ship detection in optical remote sensing images based on deep convolutional neural networks. Journal of Applied Remote Sensing. 11(4), 042611.

[87] Yang, Y., Zhuang, Y., Bi, F., et al., 2017. M-FCN: Effective fully convolutional network-based airplane detection framework. IEEE Geoscience and Remote Sensing Letters. 14(8), 1293-1297.

[88] Yang, J., Zhu, Y., Jiang, B., et al., 2018. Aircraft detection in remote sensing images based on a deep residual network and super-vector coding. Remote Sensing Letters. 9(3), 228-236.

[89] Xu, Z., Xu, X., Wang, L., et al., 2017. Deformable convnet with aspect ratio constrained NMS for object detection in remote sensing imagery. Remote Sensing. 9(12), 1312.

[90] Tang, T., Zhou, S., Deng, Z., et al., 2017. Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining. Sensors. 17(2), 336.

[91] Lin, H., Shi, Z., Zou, Z., 2017. Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images. IEEE Geoscience and Remote Sensing Letters. 14(10), 1665-1669.

[92] Liu, L., Pan, Z., Lei, B., 2017. Learning a rotation invariant detector with rotatable bounding box. arXiv:1711.09405.
DOI: https://doi.org/10.48550/arXiv.1711.09405

[93] Liu, W., Ma, L., Chen, H., 2018. Arbitrary-oriented ship detection framework in optical remote-sensing images. IEEE Geoscience and Remote Sensing Letters. 15(6), 937-941.

[94] Tang, T., Zhou, S., Deng, Z., et al., 2017. Arbitrary-oriented vehicle detection in aerial imagery with single convolutional neural networks. Remote Sensing. 9(11), 1170.

[95] Yu, Y., Guan, H., Ji, Z., 2015. Rotation-invariant object detection in high-resolution satellite imagery using superpixel-based deep Hough forests. IEEE Geoscience and Remote Sensing Letters. 12(11), 2183-2187.

[96] Zou, Z., Shi, Z., 2016. Ship detection in space-

borne optical image with SVD networks. IEEE Transactions on Geoscience and Remote Sensing. 54(10), 5832-5845.

[97] Liu, W., Anguelov, D., Erhan, D., et al., 2016. SSD: Single shot multibox detector. arXiv:1512.02325.
DOI: https://doi.org/10.48550/arXiv.1512.02325

[98] Zhong, J., Lei, T., Yao, G., 2017. Robust vehicle detection in aerial images based on cascaded convolutional neural networks. Sensors. 17(12), 2720.

[99] Ma, J., Chen, J., Ng, M., et al., 2021. Loss odyssey in medical image segmentation. Medical Image Analysis. 71, 102035.

[100] Yeung, M., Sala, E., Schönlieb, C.B., et al., 2021. A mixed focal loss function for handling class imbalanced medical image segmentation. arXiv:2102.04525.
DOI: https://doi.org/10.48550/arXiv.2102.04525

[101] Ayush, K., Uzkent, B., Tanmay, K., et al., 2020. Efficient poverty mapping using deep reinforcement learning. arXiv:2006.04224.
DOI: https://doi.org/10.48550/arXiv.2006.04224

[102] Zoph, B., Le, Q.V., 2016. Neural architecture search with reinforcement learning. arXiv: 1611.01578.
DOI: https://doi.org/10.48550/arXiv.1611.01578

[103] Tian, Y., Krishnan, D., Isola, P., 2019. Contrastive representation distillation. arXiv: 1910.10699.
DOI: https://doi.org/10.48550/arXiv.1910.10699

[104] Guerra, L., Zhuang, B., Reid, I., et al. (editors), 2020. Automatic pruning for quantized neural networks. 2021 Digital Image Computing: Techniques and Applications (DICTA); 2021 Nov 29-Dec 1; Gold Coast, Australia; USA: IEEE. p. 1-8.

[105] Yaguchi, A., Suzuki, T., Nitta, S., et al., 2019. Decomposable-Net: Scalable low-rank compression for neural networks. arXiv:1910.13141.
DOI: https://doi.org/10.48550/arXiv.1910.13141

[106] Zhou, D., Jin, X., Hou, Q., et al., 2019. Neural epitome search for architecture-agnostic network compression. arXiv:1907.05642.
DOI: https://doi.org/10.48550/arXiv.1907.05642

[107] Heo, B., Lee, M., Yun, S., et al., 2019. Knowledge distillation with adversarial samples supporting decision boundary. Proceedings of the AAAI Conference on Artificial Intelligence. 33(1), 3771-3778.

[108] Xu, Z., Hsu, Y.C., Huang, J., 2017. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. arXiv:1709.00513.
DOI: https://doi.org/10.48550/arXiv.1709.00513

[109] Heo, B., Lee, M., Yun, S., et al., 2018. Knowledge Distillation with Adversarial Samples Supporting Decision Boundary [Internet]. Available from: https://arxiv.org/pdf/1805.05532.pdf

[110] Xian, Y., Schiele, B., Akata, Z., 2017. Zero-shot learning-the good, the bad and the ugly. arXiv:1703.04394.
DOI: https://doi.org/10.48550/arXiv.1703.04394

[111] Gui, R., Xu, X., Wang, L., et al., 2018. A generalized zero-shot learning framework for PolSAR land cover classification. Remote Sensing. 10(8), 1307.

[112] Haddeler, G., Aybakan, A., Akay, M.C., et al., 2020. Evaluation of 3D LiDAR sensor setup for heterogeneous robot team. Journal of Intelligent & Robotic Systems. 100(2), 689-709.

[113] Zhou, L., Gu, X., Cheng, S., et al., 2020. Analysis of plant height changes of lodged maize using UAV-LiDAR data. Agriculture. 10(5), 146.

[114] Raj, R., Kar, S., Nandan, R., et al., 2020. Precision agriculture and unmanned aerial vehicles (UAVs). Unmanned aerial vehicle: Applications in agriculture and environment. Springer: Berlin. pp. 7-23.

[115] Wang, D., Cao, W., Zhang, F., et al., 2022. A review of deep learning in multiscale agricultural sensing. Remote Sensing. 14(3), 559.

[116] Rasti, S., Bleakley, C.J., Silvestre, G., et al., 2021. Crop growth stage estimation prior to canopy closure using deep learning algorithms.

Neural Computing and Applications. 33(5), 1733-1743.

[117] Sagan, V., Maimaitijiang, M., Bhadra, S., et al., 2021. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. ISPRS Journal of Photogrammetry and Remote Sensing. 174, 265-285.

[118] Li, X.F., Liu, B., Zheng, G., et al., 2020. Deep-learning-based information mining from ocean remote-sensing imagery. National Science Review. 7(10), 1584-1605.

[119] Kruk, R., Fuller, M.C., Komarov, A.S., et al., 2020. Proof of concept for sea ice stage of development classification using deep learning. Remote Sensing. 12(15), 2486.

[120] Han, Y., Liu, Y., Hong, Z., et al., 2021. Sea ice image classification based on heterogeneous data fusion and deep learning. Remote Sensing. 13(4), 592.

[121] Ren, Y., Li, X., Yang, X., et al., 2021. Development of a dual-attention U-Net model for sea ice and open water classification on SAR images. IEEE Geoscience and Remote Sensing Letters. 19, 1-5.

[122] Zhang, J., Zhang, W., Hu, Y., et al., 2022. An improved sea ice classification algorithm with Gaofen-3 dual-polarization SAR data based on deep convolutional neural networks. Remote Sensing. 14(4), 906.

[123] Alsharay, N.M., Chen, Y., Dobre, O.A., et al., 2022. Improved sea-ice identification using semantic segmentation with raindrop removal. IEEE Access. 10, 21599-21607.

[124] Kharazi, B.A., Behzadan, A.H., 2021. Flood depth mapping in street photos with image processing and deep neural networks. Computers, Environment and Urban Systems. 88, 101628.

[125] Singh, S., Ghosh, S., Maity, A., et al., 2022. DisasterNet: A multi-label disaster aftermath image classification model. ICT systems and sustainability. Springer: Singapore. pp. 481-490.

[126] Bickler, S.H., 2021. Machine learning arrives in archaeology. Advances in Archaeological Practice. 9(2), 186-191.

[127] Chatterjee, R., Chatterjee, A., Halder, R. (editors), 2021. Impact of deep learning on arts and archaeology: An image classification point of view. Proceedings of International Conference on Machine Intelligence and Data Science Applications; 2021 Aug; India. p. 801-810.

[128] Agapiou, A., Vionis, A., Papantoniou, G., 2021. Detection of archaeological surface ceramics using deep learning image-based methods and very high-resolution UAV imageries. Land. 10(12), 1365.

[129] Banasiak, P.Z., Berezowski, P.L., Zapłata, R., et al., 2022. Semantic segmentation (U-Net) of archaeological features in airborne laser scanning—Example of the Białowieża Forest. Remote Sensing. 14(4), 995.

[130] Yu, H., Ma, Y., Wang, L., et al. (editors), 2017. A landslide intelligent detection method based on CNN and RSG_R. 2017 IEEE International Conference on Mechatronics and Automation (ICMA); 2017 Aug 6-9; Takamatsu, Japan. USA: IEEE. p. 40-44.

[131] Ding, A., Zhang, Q., Zhou, X., et al. (editors), 2016. Automatic recognition of landslide based on CNN and texture change detection. 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC); 2016 Nov 11-13; Wuhan, China. USA: IEEE. p. 444-448.

[132] Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., et al., 2019. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sensing. 11(2), 196.

[133] Lu, H., Ma, L., Fu, X., et al., 2020. Landslides information extraction using object-oriented image analysis paradigm based on deep learning and transfer learning. Remote Sensing. 12(5), 752.

[134] Valade, S., Ley, A., Massimetti, F., et al., 2019. Towards global volcano monitoring using multisensor sentinel missions and artificial in-

telligence: The MOUNTS monitoring system. Remote Sensing. 11(13), 1528.

[135] Bountos, N.I., Papoutsis, I., Michail, D., et al., 2021. Self-supervised contrastive learning for volcanic unrest detection. IEEE Geoscience and Remote Sensing Letters. 19, 1-5.

[136] Kashinath, K., Mudigonda, M., Kim, S., et al., 2020. ClimateNet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. Geoscientific Model Development. 14(1), 107-124.

[137] Ray, A., Chakraborty, T., Ghosh, D., 2021. Optimized ensemble deep learning framework for scalable forecasting of dynamics containing extreme events. arXiv:2106.08968.

[138] Reiersen, G., Dao, D., Lütjens, B., et al., 2022. ReforesTree: A dataset for estimating tropical forest carbon stock with deep learning and aerial imagery. arXiv:2201.11192.
DOI: https://doi.org/10.1063/5.0074213

[139] Zhao, F., Sun, R., Zhong, L., et al., 2022. Monthly mapping of forest harvesting using dense time series Sentinel-1 SAR imagery and deep learning. Remote Sensing of Environment. 269, 112822.

[140] Sadiq, R., Akhtar, Z., Imran, M., et al., 2022. Integrating remote sensing and social sensing for flood mapping. Remote Sensing Applications: Society and Environment. 25, 100697.