Multispectral Semantic Segmentation for Land Cover Classification: An Overview

Leo Thomas Ramos, *Member, IEEE*, and Angel D. Sappa, *Senior Member, IEEE*

***Abstract*—Land cover classification (LCC) is a process used to categorize the earth's surface into distinct land types. This classification is vital for environmental conservation, urban planning, agricultural management, and climate change research, providing essential data for sustainable decision making. The use of multispectral imaging (MSI), which captures data beyond the visible spectrum, has emerged as one of the most utilized image modalities for addressing this task. In addition, semantic segmentation techniques play a vital role in this domain, enabling the precise delineation and labeling of land cover classes within imagery. The integration of these three concepts has given rise to an intriguing and ever-evolving research field, witnessing continuous advancements aimed at enhancing multispectral semantic segmentation (MSSS) methods for LCC. Given the dynamic nature of this field, there is a need for a thorough examination of the latest trends and advancements to understand its evolving landscape. Therefore, this article presents a review of current aspects in the field of MSSS for LCC, addressing the following key points: 1) prevalent datasets and data acquisition methods; 2) preprocessing methods for managing MSI data; 3) typical metrics and evaluation criteria used for assessing performance of methods; 4) current techniques and methodologies employed; and 5) spectral bands beyond the visible spectrum commonly utilized. Through this analysis, our objective is to provide valuable insights into the current state of MSSS for LCC, contributing to the ongoing development and understanding of this dynamic field while also providing perspectives for future research directions.**

***Index Terms*—Computer vision (CV), deep learning (DL), image segmentation, land cover classification (LCC), multispectral imaging (MSI), semantic segmentation, remote sensing, satellite imagery.**

I. Introduction

LAND cover classification (LCC) involves creating a schematic representation of the earth's surface [1]. This entails a physical and biological characterization of a specific area, which may include various elements such as forested areas, bodies of water, agricultural lands, and human constructions like cities and roads [2]. This serves not only to describe the distribution of elements on a surface but also to analyze the interaction between biological, geological, climatic, and human processes that have shaped this surface and how it has evolved over time [3]. For these reasons, LCC is crucial for various disciplines, such as ecology, geography, and climatology [4], as it provides information for managing natural resources, planning urban and rural development, and understanding changes in ecosystems due to human activity or natural phenomena [5], [6].

Traditionally, LCC was based on direct observations and manual analysis, which was laborious and prone to errors [7]. The introduction of artificial intelligence (AI) techniques, particularly computer vision (CV), has allowed for the development of semiautomatic and automatic methods, such as semantic segmentation, that achieve greater precision and efficiency than their counterparts [8], [9]. These methods are capable of identifying complex patterns and specific features, such as textures, and colors in diverse environments [10], [11]. They are trained with large quantities of labeled images [12], where each pixel of the image has been previously classified by experts [13], and learn from these examples to identify the distinctive characteristics of each type of land cover. Once trained, the models can be applied to new images, classifying each pixel according to the learned categories, resulting in an accurate and detailed land cover map.

One of the most utilized techniques for LCC is semantic segmentation. At its core, semantic segmentation involves partitioning an image into several segments [14], aiming to simplify its representation into something more meaningful and easier to analyze [14], [15]. This process relies on a variety of algorithms that analyze and classify each pixel of an image into a specific category based on the object it belongs to. As a result, the outcome is a segmented image where each pixel is tagged with a label [16], thereby transforming raw visual input into a structured and interpretable format.

While this technique has proven highly effective in LCC, it faces challenges particularly because the aerial perspective used, such as those from drones or satellites [17], covers large areas densely packed with details. This makes it difficult to differentiate between similar structures such as dense wooded areas and croplands. Moreover, the resolution of these images can be limited [18], hindering the precise identification of smaller or subtle features in the terrain. In addition, the traditional images used in this process are often of the red, green, and blue (RGB)

Manuscript received 9 May 2024; revised 15 July 2024; accepted 30 July 2024. Date of publication 5 August 2024; date of current version 26 August 2024. This work was supported in part by the Air Force Office of Scientific Research Under Award FA9550-24-1-0206, in part by the Agencia Estatal de Investigación, Ministerio de Ciencia e Innovación, under Grant PID2021-128945NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe," in part by the ESPOL project CIDIS-003-2024, in part by the Generalitat de Catalunya CERCA Program to CVCs general activities, and the Departament de Recerca i Universitats from Generalitat de Catalunya under Grant 2021SGR01499. *(Corresponding authors: Leo Thomas Ramos; Angel D. Sappa)*

Leo Thomas Ramos is with the Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain, and also with Kauel Inc., Menlo Park, Silicon Valley, CA 94025 USA (e-mail: ltramos@cvc.uab.cat, leo.ramos@kauel.com).

Angel D. Sappa is with the Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain, and also with ESPOL Polytechnic University, Guayaquil 090112, Ecuador (e-mail: asappa@cvc.uab.cat, asappa@espol.edu.ec).

Digital Object Identifier 10.1109/JSTARS.2024.3438620

type [16], [19], which represent the standard format for capturing and displaying visual information. RGB images are favored for their straightforwardness and the immediate availability of data that they offer. However, there are scenarios where the conventional RGB imagery may not suffice [20], [21]. These limitations become apparent in applications requiring the detection of features or details that are not readily distinguishable in the visible spectrum. In such cases, the reliance solely on RGB images can lead to inadequate or incomplete analyses [20], [22], [23], prompting the exploration of alternative imaging techniques.

In this scenario, multispectral imagery (MSI) stands out, offering an expansive and nuanced view of the terrestrial surface. MSI involves the collection of image data at specific wavelengths across the electromagnetic spectrum, including both the visible light range and beyond, into infrared, near-infrared (NIR), red edge (RE), and sometimes ultraviolet ranges [24]. This improves the ability to differentiate various land features, given the sensitivity of specific spectral bands to different aspects of the terrain, allowing the accurate identification of particular objects and structures [25]. For instance, the NIR band's responsiveness to vegetation health significantly aids in pinpointing areas of vigorous vegetation, water stress, or the seasonal variation in plant growth [26]. Similarly, the mid-infrared band is adept at detecting soil composition changes or identifying unique substances, such as asphalt in urban settings [27], [28]. In addition, the short-wave infrared (SWIR) serves to identify minerals, analyze vegetation, and evaluate water content [29], [30].

The integration of semantic segmentation with MSI heralds a significant leap forward in LCC [31]. By leveraging the detailed spectral information provided by MSI, semantic segmentation algorithms gain the capability to discern and categorize land features with unprecedented accuracy [23]. This synergy allows for the identification of subtle differences in the terrain that RGB imagery alone cannot capture [32]. This not only mitigates the challenges posed by the limitations of RGB images but also significantly expands the potential for comprehensive environmental analysis [31]. Consequently, the combined approach of multispectral semantic segmentation (MSSS) opens up new avenues for more detailed and accurate land cover mapping, facilitating a deeper understanding of land use patterns and the spatial dynamics of various cover types, supporting diverse applications in planning, development, and conservation efforts.

Based on the above, this work endeavors to explore the landscape of MSSS in LCC, particularly emphasizing the latest contributions over recent years. The main objective of this work is to compile a comprehensive overview of the state of research, identifying the range of methodologies, applications, and other aspects that characterize the use of MSSS for LCC. Through this effort, we intend to contribute to the ongoing development in the field, providing insights into its development and future directions. Consequently, we hope to illuminate the potentialities and challenges, fostering further exploration and innovation in this domain.

II. Related Work

Given the particular relevance of LCC using MSI images, several reviews have been conducted in the literature that analyze various aspects related to our area of study. These reviews provide valuable insights and highlight different methodologies and approaches that have been employed over the years. Hossain and Chen [33] conducted a review on geographic object-based image analysis (GEOBIA) techniques using remote sensing images. They explored various segmentation methods, detailing their application to GEOBIA, and provided an in-depth analysis of each technique's conceptual framework, along with their respective advantages and limitations. In addition, their review included a summary of the available tools and software packages used for segmentation. A significant challenge they identified in image segmentation is the selection of optimal parameters and algorithms that can create image objects corresponding to meaningful geographic entities. Furthermore, they noted an increased use of imagery beyond the traditional RGB, with several reviewed works incorporating synthetic aperture radar (SAR), panchromatic, and nonvisible channels such as NIR. Alem and Kumar [34] conducted a review focusing on deep learning (DL) methods used for land cover and land use classification. Their review highlights that convolutional neural networks (CNNs) are the preferred techniques for this task. However, they also identify the use of recurrent neural networks (RNNs) and generative adversarial networks (GANs) in developing frameworks. The review emphasizes that the most commonly used images for these tasks are from remote sensing, featuring several notable datasets. While some of these datasets are in RGB format, others are derived from satellites like Landsat, which provide multispectral (MS) capabilities. Thasveen and Suresh [35] conducted a review on land use and land cover methods. In their review, they compile various approaches, including statistical methods, machine learning (ML), and DL techniques. They highlight the advancements and increased availability of remote sensing MS data thanks to different satellite programs, such as Landsat, IKONOS, SPOT, and GeoEye. Regarding methods, despite the consistent use of CNNs and artificial neural networks, there is also a significant use of ML methods such as support vector machines (SVMs) and random forests (RFs). Digra et al. [36] conducted a review on land cover and land use classification using remote sensing images and DL. Their review emphasizes that, while DL methods are efficient and can manage the complexities of land cover and land use classification, the continuous increase in available data has introduced new layers of complexity that only DL can effectively address. In light of this, they highlight the use of CNN architectures, such as ResNet, DenseNet, and GoogleNet, for this task. They also mention various data sources, particularly satellites like Landsat, Sentinel, and MODIS, which provide MS data. In addition, the review identifies commonly used software tools for this task. Despite the valuable insights provided by these reviews, there remains a need for a comprehensive analysis that focuses specifically on MSI for LCC. While previous reviews have addressed various techniques and advancements, our review uniquely centers on summarizing current contributions and methodologies specific to MSSS. By providing an in-depth overview of the latest developments and applications in this field, we aim to offer a valuable resource that enhances understanding and supports further advancements in LCC using MS data.

III. Methods

To conduct this review, we adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statements [37], the most commonly used reporting guidelines for systematic reviews. Building on this foundation, we have formulated specific research questions and established a set of inclusion and exclusion criteria to guide our selection process. The methodology employed in conducting this review will be outlined in the following sections.

A. Research Questions

This systematic review aims to thoroughly investigate the field of MSSS for LCC by addressing the following research questions.

Q1 What are the most commonly used datasets and data acquisition methods in the field of MSSS for LCC?

Q2 What specific preprocessing methods or techniques are employed to handle MSI data?

Q3 Which metrics and evaluation criteria are most commonly used in assessing the performance of techniques and methods?

Q4 What techniques and methodologies are primarily used in MSSS for LCC?

Q5 Which spectral bands beyond the visible spectrum are most commonly utilized in MSSS for LCC?

B. Eligibility Criteria

We adhered to specific inclusion and exclusion criteria to ensure the relevance and quality of the articles selected for review. For inclusion, articles were considered eligible if they satisfied the following conditions: (IC1) empirical research focused specifically on MSSS for LCC; (IC2) research published between 2020 and 2024 in peer-reviewed journals or presented at significant conferences; and (IC3) research utilizing MSI up to 36 bands. On the other hand, exclusion criteria were applied to articles that failed to meet any of the following requirements: (EC1) review articles, meta-analyses, editorials, commentaries, and other forms of secondary research; and (EC2) non-English research articles.

C. Search Strategy

To gather the articles for our review, we crafted a query that contains a selection of keywords that cover the study domain. The search string used was: (“semantic segmentation” OR “segmentation” OR “image segmentation”) AND (“multispectral” OR “multispectral imaging” OR “multispectral imagery”) AND (“land cover” OR “land cover mapping” OR “land cover classification” OR “land cover analysis”). This query was applied to the metadata of articles, including titles, abstracts, and keywords, to ensure the comprehensive coverage of pertinent studies. In addition, we applied filters to refine the search results further, ensuring the retrieval of the most relevant articles. We utilized a variety of databases and search engines for our literature search, including IEEE Xplore,[1] ScienceDirect,[2] Springer,[3] ACM,[4] and Taylor & Francis.[5] We opted not to include other sources such as trial registers or grey literature.

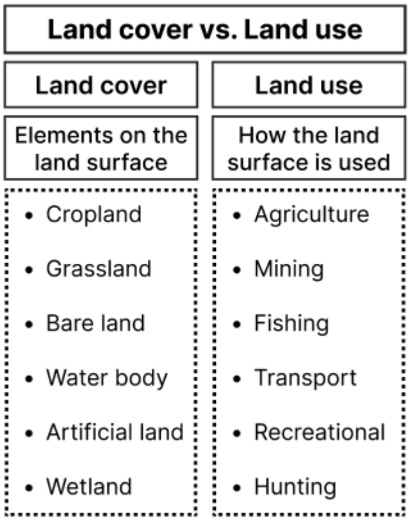


Fig. 1. Difference between land cover and land use.

IV. Brief Fundamentals

In this section, we present a concise overview of the fundamental concepts underpinning our field of study: LCC, semantic segmentation, and MSI. This primer is designed to establish a shared understanding of the key principles and terminology that are crucial for exploring the advanced methodologies discussed later in this review.

A. Land Cover Classification

It is worth remembering that LCC refers to the process of categorizing the earth’s surface into distinct classes. This process involves examining an image to identify different objects and assigning them to specific categories. This transforms complex visual information into clearly defined segments, generating an organized map that describes the distribution of different types of land cover. In this field, it is important to differentiate between land cover and land use, two terms that are often confused. While both terms relate to the earth’s surface, each captures a distinct dimension. As shown in Fig. 1, land cover refers to the physical covering of the land, describing the natural and artificial elements that compose it [2], [38]. This includes vegetation, artificial surfaces, water bodies, and barren lands. In contrast, land use focuses on human usage of the land [39], [40], highlighting the activities and functions that take place on it [38], such as agriculture, recreation, and commerce. In summary, land cover and land use are interrelated but distinct concepts that provide complementary perspectives on the landscape.

[1] [Online]. Available: https://ieeexplore.ieee.org
[2] [Online]. Available: www.sciencedirect.com
[3] [Online]. Available: www.link.springer.com
[4] [Online]. Available: https://dl.acm.org/
[5] [Online]. Available: https://www.tandfonline.com

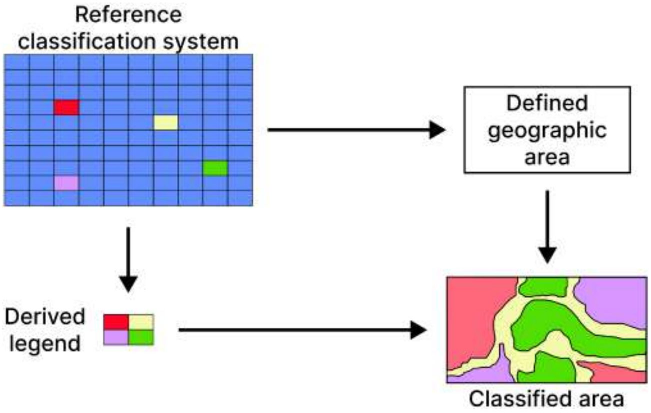
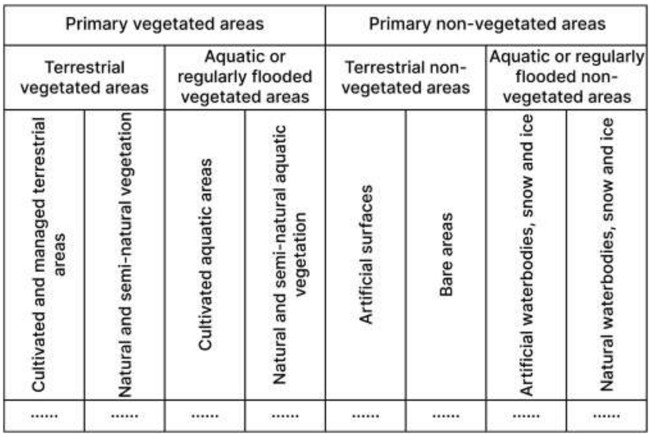


Fig. 2. Example of an LCC process of an area using a reference system.



Primary vegetated areas				Primary non-vegetated areas			
Terrestrial vegetated areas		Aquatic or regularly flooded vegetated areas		Terrestrial non-vegetated areas		Aquatic or regularly flooded non-vegetated areas	
Cultivated and managed terrestrial areas	Natural and semi-natural vegetation	Cultivated aquatic areas	Natural and semi-natural aquatic vegetation	Artificial surfaces	Bare areas	Artificial waterbodies, snow and ice	Natural waterbodies, snow and ice
......	......	......	......	......	......	......	......

Fig. 3. Upper level categories of the LCCS proposed by the FAO.

Prior to the categorization of any scene, establishing the classes is one of the initial steps in LCC [41]. While it is possible for researchers to define their categories based on the specific needs of their study or regional characteristics, there are structured systems that provide guidelines for defining classes and levels of classification. These standards ensure coherence and are particularly useful for comparative analyses across various locations or different time periods, maintaining consistency across studies. Fig. 2 shows an example of how a reference classification system is used for mapping an area. One of the most renowned guides is the Land Cover Classification System (LCCS)[6] established by the Food and Agriculture Organization (FAO). This system proposes a hierarchical classification that starts with broad levels, which then lead to more specific subdivisions [42], as seen in Fig. 3. This structure allows users to accurately identify and describe a wide range of land cover types, from forests and grasslands to urban areas, croplands, and bare surfaces. This broad spectrum greatly facilitates the conceptualization of studies where the proper establishment of classes is not trivial [42], [43]. Typically, many studies utilize this classification system, adopting specific levels tailored to their unique case requirements.

Within the methods applied for LCC, these can broadly be divided into two categories: pixel-based image analysis (PBIA) and object-based image analysis (OBIA) [44], [45]. The PBIA approach classifies each pixel individually, based on its spectral signature, without considering the context or spatial relationships with surrounding pixels [45], [46]. This can be effective in scenarios where there is high homogeneity within the land cover classes. However, this technique may face limitations in areas where land cover types exhibit similar spectral characteristics or in environments where pixel mixing is frequent [47]. On the other hand, OBIA groups pixels into larger objects based on similarity and other criteria, such as texture, shape, and proximity. This involves an initial segmentation step where pixels are grouped into larger objects [48], which are then classified as units [45]. In addition, OBIA can utilize segmentations at multiple scales [48], [49], [50], capturing various levels of detail. This approach allows the method to effectively identify broad homogeneous areas as well as finer more heterogeneous regions. This results in a more contextual and detailed interpretation, proving especially useful in heterogeneous landscapes where different types of land cover interact in complex ways [48]. Fig. 4 shows the comparison between PBIA and OBIA graphically.

With the technological advancements in the field of AI, DL techniques have increasingly been incorporated into LCC approaches. Specifically, methods such as semantic segmentation have become pivotal in enhancing the accuracy and efficiency of classifying land cover. These models leverage complex computational algorithms to analyze and interpret the vast amounts of image data more effectively than traditional methods. This evolution marks a significant shift toward more sophisticated automated analysis in environmental monitoring and land management.

B. Semantic Segmentation

Image segmentation is one of the fundamental tasks in CV. It is defined as the process of partitioning a digital image into multiple segments, transforming it into something that is more meaningful and easier to analyze [51]. This involves a pixel-level analysis where each pixel in an image is classified or assigned to a specific category [52]. This results in the formation of clusters where all pixels in the same group share certain characteristics [53]. This technique not only allows for locating and identifying a particular object but also provides information about the size, shape, and even orientation [54], [55], making it more comprehensive than mere classification and detection.

Broadly, image segmentation can be categorized into two types: instance segmentation and semantic segmentation [56]. Semantic segmentation involves the process of assigning a label to every pixel in an image such that pixels that are part of the same object type share the same label [51]. This type of segmentation does not differentiate between different instances of the same object [57]; instead, it aims to categorize all pixels of similar types into the same segment across the whole image. It is particularly useful for understanding the layout of an environment

[6] [Online]. Available: https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1036361/

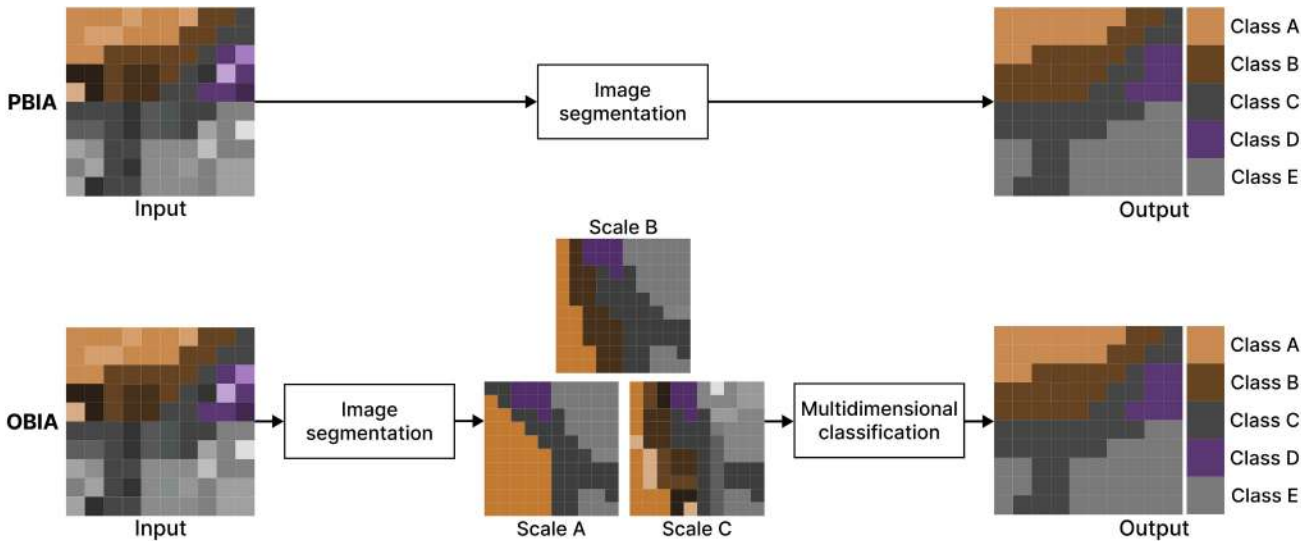


Fig. 4. Comparison between PBIA and OBIA approaches.

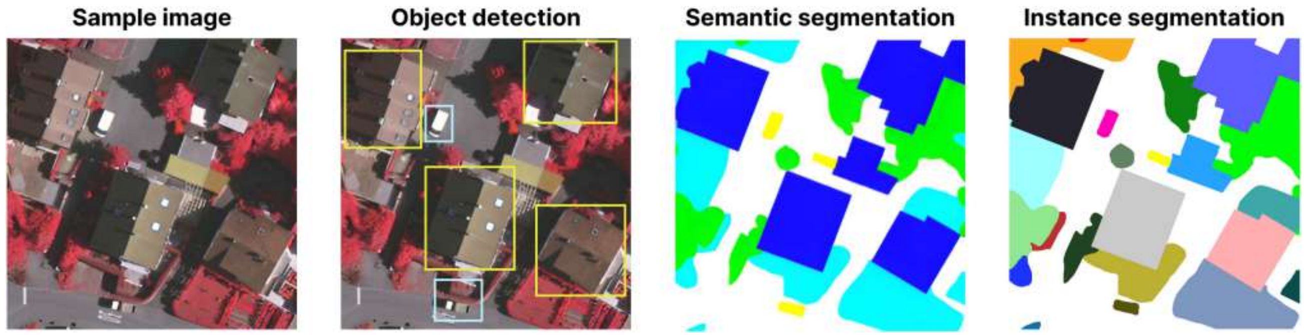


Fig. 5. Comparison between different CV tasks.

by labeling areas such as roads, buildings, cars, and trees [58], which helps in tasks such as autonomous driving and LCC. Instance segmentation goes a step further by not only segmenting the image into defined categories but also distinguishing between different instances of the same category [59], [60]. For example, if there are several cars in an image, instance segmentation will identify and separate each specific car rather than marking them all as one category. This is especially beneficial in scenarios where the identification of individual items is required, such as in object tracking and counting [61]. Both types of segmentation provide valuable insights for analyzing complex images, but their application depends on the specific needs of the task at hand, whether understanding a scene as a whole (semantic) or identifying and differentiating individual elements within the scene (instance). Fig. 5 shows the difference between this tasks.

Regarding segmentation methods, traditional techniques, such as thresholding, region-based segmentation, and edge detection, have laid the groundwork for this field [62]. Thresholding simplifies images to binary levels based on pixel values [62], [63]. Region-based algorithms group pixels into larger areas based on similar statistical characteristics like color and texture, utilizing techniques like region growing and split-and-merge [64]. Edge detection identifies object boundaries by detecting discontinuities in pixel intensity [65], using operators such as Sobel or Canny [65], [66]. As the field of CV has advanced, segmentation methods have evolved from simple traditional techniques to more complex approaches based on ML and DL [16]. Following initial techniques, such as thresholding, ML methods began to play a crucial role in improving image segmentation. Techniques such as SVM [67], K-means, decision tree [68], and RF [69] are representative of this group [70]. These methods use manually extracted features to train models capable of discerning between different types of segments based on patterns learned from the data [71], [72]. These approaches provide better management of variations and complexities in images than traditional methods, facilitating more precise segmentation tailored to the specific characteristics of each context.

The introduction of DL and neural networks has further revolutionized the field of semantic segmentation [73]. DL models offer significant advantages over ML techniques as they are capable of automatically learning hierarchical feature representations from large amounts of data [74], thereby eliminating the

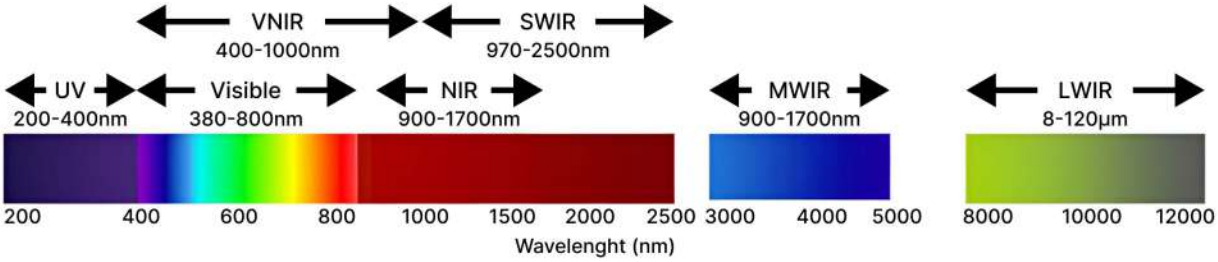


Fig. 6. Example of the electromagnetic spectrum and its divisions into spectral bands. Band ranges can vary with the imaging equipment used, making this figure a general depiction for segmenting the spectrum in various applications, not an exact standard for all imaging systems.

need for manual feature extraction [75], which is often labor-intensive and prone to errors. Specifically, CNNs are particularly well suited for image processing tasks [76]. They utilize layers of convolutions, which apply filters to an input image to create feature maps that summarize the presence of specific features in the input [77], [78]. This way, early layers might detect edges or textures, while deeper layers might identify more complex patterns such as parts of objects [77], [79], [80]. This hierarchical approach to learning features makes CNNs incredibly effective for tasks that require understanding of visual data [81], as each layer builds on the information processed by the previous one, leading to a detailed and comprehensive understanding of the image content. These CNN approaches have proven to be particularly effective in handling complex image segmentation tasks [73], demonstrating substantial improvements in accuracy, generalization, and the ability to handle diverse and challenging environments. Within this group, architectures such as UNet [82], fully convolutional network [83], and SegNet [84] are some of the pioneers in demonstrating the capability of CNN methods, becoming some of the most recognized and used in semantic image segmentation [62], [85].

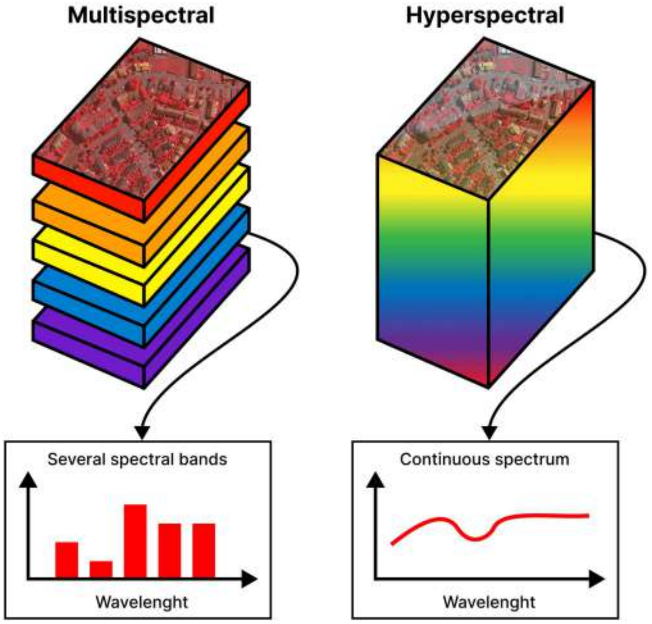


Fig. 7. Comparative diagram between MSI and HSI.

C. Multispectral Imaging

MSI involves capturing data at multiple wavelengths across the electromagnetic spectrum, beyond the narrow band of visible light [24], [86]. This method collects information from various spectral bands, not only including the standard RGB that human eyes perceive but also extending into invisible wavelengths [86]. Each spectral band is capable of detecting unique attributes of surface materials, revealing details that are not discernible through conventional imaging.

The capture of MSI requires specialized devices and sensors designed to detect and record data across various wavelengths [87]. These sensors are often part of sophisticated imaging systems that can include stand-alone cameras, satellite-mounted sensors, and airborne systems [88], each tailored for specific applications. Stand-alone MS cameras are commonly used in handheld or drone-mounted configurations, offering flexibility for ground-level or aerial surveys [89], [90]. Satellite sensors provide broader coverage, making them ideal for environmental monitoring and geographic mapping on a global scale [91], [92]. In addition, airborne systems, mounted on aircraft, bridge the gap between satellite- and ground-based observations by providing detailed imagery over large areas, which is essential for precision agriculture, forestry, and land management [86].

In this regard, it is important to recognize and understand the differences between MS and hyperspectral images, two types of similar images but with distinct capabilities and applications in the capture and analysis of spectral data. As mentioned earlier, MSI captures information in bands beyond the visible spectrum; however, these typically focus on broader ranges, wider bandwidth, and finer resolution [93], making them especially useful for general applications, such as agriculture, environmental monitoring, and cartography. Fig. 6 presents an illustrative example of the electromagnetic spectrum and its divisions into various spectral bands. On the other hand, hyperspectral imaging (HSI) also captures data beyond the visible spectrum, but it collects detailed information at an almost continuous level along the spectrum with a narrow bandwidth [93], [94], as illustrated in Fig. 7. This allows for a detailed characterization of each point in the image [95]. This granularity in data collection enables HSI to identify the subtleties of materials with high precision, which is useful in applications that require precise identification of certain elements, such as in geology [96], and detection of chemical pollutants [97]. However, this type of imaging comes with its challenges. Unlike MSI, which, due to the fewer number of bands, are easier to process and analyze, HSI requires greater computational capacity and more sophisticated processing techniques. Likewise, their capture demands more sophisticated equipment that is often much more expensive than that required to capture MSI [98], [99], which represents a significant barrier in terms of initial investment and maintenance for many research projects and commercial applications.

An important point to note is that there is no universally accepted standard that precisely defines the boundary between MSI and HSI in terms of the number of bands. While MSI is typically limited to a few bands, HSI may contain hundreds of spectral bands. The distinction is usually based more on the application and analysis intended to be performed. Some research [86] suggests that MSI contains three to seven spectral bands, while HSI ranges from 10 to 100 bands. Others [100] state that MSI data comprise between five and ten bands, while HSI includes between 100 and 200. Meanwhile, the United States Geological Survey (USGS)[7] specifies that MSI can have up to 36 bands, whereas HSI can contain hundreds and even thousands. Therefore, although no fixed limit is established, for a general criterion, it can be stated that MSI contains up to a few tens of bands, while HSI contains hundreds. For the purposes of this review, the criterion established by the USGS will be considered.

In sum, MSI provides a valuable tool for earth observation and analysis, combining an effective balance between spectral resolution, spatial coverage, and operational simplicity. Although HSI offers greater depth in spectral analysis, the complexity and costs associated with its use limit its applicability in many contexts. For these reasons, this review will focus on MSI, capitalizing on its ability to deliver detailed and relevant information without the technical and financial demands of HSI. The choice of MSI enables broader access to remote sensing technology, allowing more practical and economical implementation across various fields of study.

In the context of MSI, spectral indices play a significant role in the interpretation and analysis of captured data. These indices are mathematical calculations made from specific spectral bands of an image [101] and are used to identify and quantify particular characteristics of the earth's surface. For example, the Normalized Difference Vegetation Index (NDVI) is one of the most well known and is calculated using the red and NIR bands [102], [103]. The NDVI is especially valuable for assessing the amount and health of vegetation in a given area, providing crucial information for agriculture, forest management, and environmental monitoring [102], [104]. Another widely used index is the Normalized Difference Water Index (NDWI), which is calculated from the NIR and green bands [105]. The NDWI is particularly useful for water resource management and flood detection, as it highlights water-saturated areas, allowing a clear distinction between aquatic and terrestrial zones [105], [106]. Including spectral indices in the analysis of MSI allows researchers to fully leverage the unique properties of different bands of the spectrum, facilitating a more detailed and specific interpretation of the data. In addition to the NDVI and NDWI, there are numerous other spectral indices that can be derived from the available bands in MSI. Each of these indices is designed to highlight specific features and address particular needs in various fields of study, such as mineralogy, soil quality, and the identification of urban areas. Table I lists some of the most commonly used indices.

In addition to spectral indices, MSI can include or be complemented with digital surface models (DSMs) and normalized digital surface models (nDSMs). These models are crucial for providing 3-D information about the earth's surface and its morphology. A DSM represents the elevation of all objects on the ground [112], including vegetation, buildings, and other structures, while the nDSM is generated by subtracting the elevation of the natural terrain, obtained from a digital terrain model [113], resulting in a model that exclusively reflects the height of features above the natural surface, such as buildings and trees. Incorporating these models into the analysis of MSI allows for a deeper and more detailed understanding of the environment being analyzed. For example, DSM and nDSM are widely used in urban planning, environmental impact studies, and natural resource management [112], [114], as they provide a precise view of the spatial distribution and height of objects within a given region. This information is invaluable for tasks such as planning new infrastructures, and managing vegetation in urban and rural areas. Moreover, combining spectral indices with altimetric data from DSM and nDSM can enrich land cover analysis, facilitating more precise segmentation and much more detailed and comprehensive studies.

V. Datasets and Image Acquisition

In this section, we detail the datasets and image acquisition methods commonly used in the field of MSSS for LCC identified in this review. We explore the variety of data sources, the specific spectral bands or range covered, and the techniques employed to gather and prepare these datasets for effective analysis.

A. Satellite Imaging Services

In our review, a significant number of studies focus on application-driven research and case studies. These works are clearly distinct from those primarily aimed at developing or testing new methodologies, and hence, the reliance on precompiled benchmark datasets is noticeably less common. Instead, these studies often require the acquisition of specific imagery tailored to their unique research objectives.

In this context, satellite imaging services emerge as one of the primary sources for obtaining such data. These services offer the flexibility to provide images across a wide range of locations and geographical conditions, diverse bands, and multitemporal captures, making them invaluable tools for a broad array of applications. Given the frequent use of these satellite services for data acquisition among the reviewed studies, the following details the most utilized ones, addressing the different characteristics and functionalities of these services.

[7] [Online]. Available: https://www.usgs.gov/

TABLE I
SOME COMMON SPECTRAL INDICES

Index	**Abbreviation**	**Formula**	**Ref.**[a]
Normalized Difference Vegetation Index	NDVI	$\mathrm{NDVI} = \frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+\mathrm{Red}}$	[102]
Normalized Difference Water Index	NDWI	$\mathrm{NDWI} = \frac{\mathrm{Green}-\mathrm{NIR}}{\mathrm{Green}+\mathrm{NIR}}$	[105]
Modified Normalized Difference Water Index	MNDWI	$\mathrm{MNDWI} = \frac{\mathrm{Green}-\mathrm{SWIR}}{\mathrm{Green}+\mathrm{SWIR}}$	[107]
Enhanced Vegetation Index	EVI	$\mathrm{EVI} = 2.5 \times \frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+6\times\mathrm{Red}-7.5\times\mathrm{Blue}+1}$	[108]
Normalized Difference Moisture Index	NDMI	$\mathrm{NDMI} = \frac{\mathrm{NIR}-\mathrm{SWIR}}{\mathrm{NIR}+\mathrm{SWIR}}$	[109]
Normalized Difference Built-up Index	NDBI	$\mathrm{NDBI} = \frac{\mathrm{SWIR}-\mathrm{NIR}}{\mathrm{SWIR}+\mathrm{NIR}}$	[110]
Soil-Adjusted Vegetation Index	SAVI	$\mathrm{SAVI} = \frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+\mathrm{Red}+0.5} \times (1.5)$	[111]
Optimized Soil Adjusted Vegetation Index	OSAVI	$\mathrm{OSAVI} = \frac{\mathrm{NIR}-\mathrm{Red}}{\mathrm{NIR}+\mathrm{Red}+16}$	[109]
Modified Soil-Adjusted Vegetation Index	MSAVI2	$\mathrm{MSAVI2} = \frac{(2\times\mathrm{NIR}+1)-\sqrt{(2\times\mathrm{NIR}+1)^2-8\times(\mathrm{NIR}-\mathrm{Red})}}{2}$	[109]

[a] "Ref." column lists sources where indices are described or applied, not necessarily the original publication of the index.

TABLE II
OPERATIONAL FEATURES OF SENTINEL MISSIONS

Mission	**Launch date**	**End of operations**	**Repeat cycle**	**Instruments**
Sentinel-1	3 April 2014	Active	12 days	Synthetic Aperture Radar
Sentinel-2 (A/B)	23 June 2015	Active	10 days	MultiSpectral Instrument
Sentinel-3	16 February 2016	Active	27 days	Sea and Land Surface Temperature Radiometer (SLSTR) + Ocean and Land Colour Instrument (OLCI) + Synthetic Aperture Radar Altimeter (SRAL) + Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS) + Microwave Radiometer (MWR) + Laser Retroreflector (LRR)
Sentinel-4	To be confirmed	–	60 minutes	Ultraviolet–Visible–Near-Infrared Hyperspectral Imaging Spectrometer
Sentinel-5P	13 October 2017	Active	17 days	TROPOspheric Monitoring Instrument (TROPOMI)
Sentinel-5	To be confirmed	–	29 days	TROPOMI
Sentinel-6	21 November 2020	Active	10 days	A Ku/C-band nadir-pointing SAR altimeter (Poseidon-4) + Advanced Microwave Radiometer for Climate (AMR-C) + High-Resolution Microwave Radiometer (HRMR)

1) Sentinel: The Sentinel[8] satellites are a fleet of earth observation equipment developed by the European Space Agency as part of the Copernicus program. Their purpose is to provide a comprehensive and precise earth observation system, delivering crucial information on various aspects of the planet that facilitates informed decision making regarding the environment.

A significant feature of the Sentinel project is its commitment to open access, as all data collected by the Sentinel satellites are available for free to anyone in need. In addition, it includes a Hub,[9] where various services for processing the data are accessible. The backbone of the project is its fleet of satellites, with currently six consolidated missions, each specifically oriented toward monitoring certain aspects of the planet, as indicated below. In addition, Table II provides the operational features of each Sentinel mission.

1) *Sentinel-1:* It primarily focuses on earth's surface and maritime surveillance, offering capabilities for monitoring environmental changes and supporting disaster management efforts.
2) *Sentinel-2:* It aims at monitoring land and vegetation, providing critical data for agriculture, forest monitoring, and environmental change detection.
3) *Sentinel-3:* It is dedicated to observing the earth's surface, oceans, and atmosphere, contributing vital information for studying sea surface temperatures, ocean ecosystems, and land vegetation.
4) *Sentinel-4 and Sentinel-5:* It is dedicated to atmospheric monitoring; these missions measure trace gases such as nitrogen dioxide, ozone, formaldehyde, sulfur dioxide,

[8] [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/home

[9] [Online]. Available: https://www.sentinel-hub.com/

TABLE III
Description of Sentinel-2 (A/B) Bands

Spectral bands	**Central wavelength** (μm)	**Resolution** (m)
Band 1 — Coastal aerosol	0.443	60
Band 2 — Blue	0.490	10
Band 3 — Green	0.560	10
Band 4 — Red	0.665	10
Band 5 — Red edge	0.705	20
Band 6 — Red edge	0.740	20
Band 7 — Red edge	0.783	20
Band 8 — NIR	0.842	10
Band 8a — NIR narrow	0.865	20
Band 9 — Water vapor	0.940	60
Band 10 — Cirrus	1.375	60
Band 11 — SWIR	1.160	20
Band 12 — SWIR	2.190	20

methane, and carbon monoxide, as well as aerosols affecting air quality. Sentinel-4 is hosted on geostationary satellites, while Sentinel-5 Precursor and Sentinel-5 are in polar orbit.

5) *Sentinel-6:* It concentrates on studying the oceans, particularly monitoring sea level rise and offering data for operational oceanography and climate research.

In this review, Sentinel-2 has been pinpointed as the primary tool employed in the domain of semantic segmentation in MSI. Its prominence is attributed to the distinctive capabilities of its MS sensors, providing researchers with precise data essential for in-depth environmental and terrestrial analysis. Its wide spectral range facilitates the extraction of critical information regarding vegetation health, moisture levels, and other vital environmental indicators, making Sentinel-2 exceptionally suited for numerous applications. Table III offers a detailed description of the Sentinel-2 bands.

2) WorldView: WorldView[10] is a series of commercial earth observation satellites owned by Maxar Technologies. These satellites are renowned for their capability to capture images with high spatial and spectral resolution. For this reason, they are a valuable source of data for various applications, including cartography, urban planning, disaster management, defense and intelligence, and environmental monitoring.

In our review, this satellite service was identified as one of the most utilized options for data acquisition. The series consists of four satellites, enhancing global monitoring and analysis capabilities. Table IV presents the operational features of each WorldView satellite, while Table V details the band configurations for each satellite, providing comprehensive insights into their data collection capabilities.

3) Landsat: The Landsat[11] program is the world's longest-running earth satellite imaging service, beginning its operations in 1972. It is a joint initiative between NASA and the USGS. Similar to the Sentinel project, Landsat is committed to free access to information, as all data collected by Landsat satellites are freely available to anyone in need through the USGS Earth Explorer portal.

To date, nine Landsat satellites have been launched, with each iteration enhancing and expanding its capabilities by integrating more advanced sensors that provide valuable data for various applications. However, not all are currently active, as several have completed their mission lifespan. Moreover, Landsat-6 never entered service due to failing to reach orbit. For a comprehensive understanding of each satellite's characteristics, Table VI provides the operational features, while Table VII specifically details the spectral bands configuration of each satellite.

4) Gaofen: Gaofen[12] satellites are part of the Chinese earth observation initiative known as the China High-Resolution Earth Observation System. This program aims to develop and deploy a series of high-resolution observation satellites to provide data for applications such as urban planning, road design, crop estimation, disaster prevention and mitigation, and environmental protection.

To date, multiple satellites within the Gaofen series have been launched. However, the level of open access to their data may vary depending on the specific satellite and the type of data involved. While some Gaofen datasets are freely available to users under certain conditions, access to others may require permissions or adherence to specific access policies. Nonetheless, Gaofen offers a significant amount of accessible data, which has made it a popular choice for data acquisition. Specifically, the Gaofen-1 and Gaofen-2 series are usually the most frequently utilized by researchers. These two versions are particularly focused on the monitoring and observation of land, providing critical data for terrestrial analysis. Table VIII displays the operational features of these three versions, while Table IX details their spectral bands.

5) Other Satellite Services: In addition to the satellite services described previously, our review also identified a number of studies that, albeit less frequently, utilize other satellite services for data acquisition, specifically Pléiades,[13] Ziyuan-3,[14] and SPOT-6/7.[15] These alternative services still play a vital role in specific research contexts, offering unique dataset and observation capabilities that complement the more commonly used satellites. To provide an overview of these additional services and their key characteristics, Table X summarizes the most important features of each of these services.

B. Benchmark Datasets

Benchmark datasets are those that have been carefully curated and are widely acknowledged within the scientific community. Unlike data extracted for specific applications and generally used only once, these datasets are designed to be reusable, providing a standardized foundation for validation and comparison of results over time and across different studies. In the field of semantic segmentation in MSI, benchmark datasets are crucial for advancing the development of the field. They are invaluable for researchers seeking to assess the effectiveness of

[10] [Online]. Available: https://earth.esa.int/eogateway/missions/worldview

[11] [Online]. Available: https://landsat.gsfc.nasa.gov/

[12] [Online]. Available: https://chinaspacereport.wordpress.com/spacecraft/gaofen/

[13] [Online]. Available: https://earth.esa.int/eogateway/missions/pleiades

[14] [Online]. Available: https://www.eoportal.org/satellite-missions/zy-3a

[15] [Online]. Available: https://www.eoportal.org/satellite-missions/spot-6-7

TABLE IV
OPERATIONAL FEATURES OF WORLDVIEW MISSIONS

Mission	Launch date	End of operations	Repeat cycle	Instruments
WorldView-1	18 September 2007	Active	14 days	Panchromatic camera
WorldView-2	08 October 2009	Active	1.1 days	Panchromatic and multispectral camera (PMC)
WorldView-3	13 August 2014	Active	1 day	PMC
WorldView-4	11 November 2016	Active	1 day	PMC

TABLE V
DESCRIPTION OF WORLDVIEW BANDS

Spectral bands	Wavelength (μm)	Resolution (m)
WorldView-1		
Panchromatic	0.45-0.90	0.50
WorldView-2		
Panchromatic	0.45-0.80	0.46
Band 1 — Coastal Blue	0.40-0.45	1.85
Band 2 — Blue	0.45-0.51	1.85
Band 3 — Green	0.51-0.58	1.85
Band 4 — Yellow	0.585-0.625	1.85
Band 5 — Red	0.63-0.69	1.85
Band 6 — Red edge	0.705-0.745	1.85
Band 7 — NIR	0.77-0.895	1.85
Band 8 — NIR	0.86-1.04	1.85
WorldView-3		
Panchromatic	0.45-0.80	0.31
Band 1 — Coastal Blue	0.40-0.45	1.24
Band 2 — Blue	0.45-0.51	1.24
Band 3 — Green	0.51-0.58	1.24
Band 4 — Yellow	0.585-0.625	1.24
Band 5 — Red	0.63-0.69	1.24
Band 6 — Red edge	0.705-0.745	1.24
Band 7 — NIR	0.77-0.895	1.24
Band 8 — NIR	0.86-1.04	1.24
Band 9 — SWIR	1.195-1.225	7.5
Band 10 — SWIR	1.550-1.590	7.5
Band 11 — SWIR	1.640-1.680	7.5
Band 12 — SWIR	1.710-1.750	7.5
Band 13 — SWIR	2.145-2.185	7.5
Band 14 — SWIR	2.185-2.225	7.5
Band 15 — SWIR	2.235-2.285	7.5
Band 16 — SWIR	2.295-2.365	7.5
WorldView-4		
Panchromatic	0.45-0.80	0.31
Band 1 — Blue	0.45-0.51	1.24
Band 2 — Green	0.51-0.58	1.24
Band 3 — Red	0.655-0.69	1.24
Band 4 — NIR	0.78-0.92	1.24

their approaches against the most advanced existing methods, as they serve as a standard against which new methodologies can be rigorously tested; this, in turn, allows for a clear measurement of progress in the field.

In the following sections, we delve into detailed descriptions of some of the most utilized benchmark datasets, as identified in our literature review. These accounts aim to provide readers an understanding of the specific features of each dataset, including their spectral bands, categorization, and the nature of their annotations, thereby providing insights into their applicability for different research objectives. In addition, Table XI provides a summary of every benchmark dataset found in this review.

1) ISPRS Vaihingen: Among the benchmark datasets identified in our review, the ISPRS Vaihingen[16] dataset stands out as the most frequently mentioned. Developed by the International Society for Photogrammetry and Remote Sensing (ISPRS) for semantic segmentation tasks focused on land covering, this dataset features images of an urban environment, specifically the town of Vaihingen, Germany. This town is characterized as a relatively small community with many detached houses and small multistory buildings.

The ISPRS Vaihingen dataset was captured using digital aerial cameras and comprises 33 patches of different sizes. Each patch contributes to a comprehensive true orthophoto (TOP) mosaic, and the collection is augmented with DSMs generated via dense image matching techniques. The ground sampling distance (GSD) for both the TOP and the DSM is 9 cm. The dataset is provided in TIFF format, featuring RG-NIR channels. It includes pixel-level annotations for six categories: impervious surface, building, low vegetation, tree, car, and clutter/background, as shown in Fig. 8. These categories represent the six most common classes encountered in land cover studies, encompassing both natural and man-made elements crucial for accurate environmental mapping and analysis.

2) ISPRS Potsdam: The second most recurrent dataset in our review is the ISPRS Potsdam.[17] Like the Vaihingen dataset, this one was also developed by the ISPRS and captures the essence of a German city, Potsdam, renowned for its historic urban landscape featuring large building blocks, narrow streets, and a dense settlement structure, as shown in Fig. 9.

The dataset shares several characteristics with the ISPRS Vaihingen as the images were captured by digital aerial cameras, are derived from a larger TOP, and include DSMs generated from dense image matching techniques. However, the GSD for both

[16] [Online]. Available: https://www.isprs.org/education/benchmarks/Urban\brkSemLab/2d-sem-label-vaihingen.aspx

[17] [Online]. Available: https://www.isprs.org/education/benchmarks/Urban\brkSemLab/2d-sem-label-potsdam.aspx

TABLE VI
OPERATIONAL FEATURES OF LANDSAT MISSIONS

Mission	**Launch date**	**End of operations**	**Repeat cycle**	**Instruments**
Landsat 1	23 July 1972	6 January 1978	18 days	MS scanner
Landsat 2	22 January 1975	25 February 1982	18 days	MS scanner
Landsat 3	5 March 1978	31 March 1983	18 days	MS scanner
Landsat 4	16 July 1982	14 December 1993	16 days	MS scanner + Thematic Mapper (TM)
Landsat 5	1 March 1984	5 June 2013	16 days	MS scanner + TM
Landsat 6	5 October 1993	5 October 1993	16 days	MS scanner + TM
Landsat 7	15 April 1999	6 April 2022	16 days	Enhanced Thematic Mapper Plus (ETM+)
Landsat 8	11 February 2013	Active	16 days	Operational Land Imager (OLI) + Thermal Infrared Sensor (TIRS)
Landsat 9	27 September 2021	Active	8 days	Operational Land Imager 2 (OLI-2) + Thermal Infrared Sensor 2 (TIRS-2)

TABLE VII
DESCRIPTION OF LANDSAT BANDS

Spectral bands	**Wavelength** (μm)	**Resolution** (m)
Landsat 1-3		
Band 4 — Green	0.5-0.6	60
Band 5 — Red	0.6-0.7	60
Band 6 — NIR	0.7-0.8	60
Band 7 — NIR	0.8-0.11	60
Landsat 4-5 (MS scanner)		
Band 1 - Green	0.5-0.6	60
Band 2 - Red	0.6-0.7	60
Band 3 - NIR	0.7-0.8	60
Band 4 - NIR	0.8-0.11	60
Landsat 4-5 (TM)		
Band 1 — Blue	0.45-0.52	30
Band 2 — Green	0.52-0.60	30
Band 3 — Red	0.63-0.69	30
Band 4 — NIR	0.76-0.90	30
Band 5 — SWIR	1.55-1.75	30
Band 6 — Thermal	10.40-12.50	120 (30)
Band 7 — SWIR	2.08-2.35	30
Landsat 7		
Band 1 — Blue	0.45-0.52	30
Band 2 — Green	0.52-0.60	30
Band 3 — Red	0.63-0.69	30
Band 4 — NIR	0.76-0.90	30
Band 5 — SWIR	1.55-1.75	30
Band 6 — Thermal	10.40-12.50	60 (30)
Band 7 — SWIR 2	2.08-2.35	30
Band 8 — Panchromatic	0.52-0.90	15
Landsat 8-9		
Band 1 — Coastal aerosol	0.43-0.45	30
Band 2 — Blue	0.45-0.51	30
Band 3 — Green	0.53-0.59	30
Band 4 — Red	0.64-0.67	30
Band 5 — NIR	0.85-0.88	30
Band 6 — SWIR	1.57-1.65	30
Band 7 — SWIR	2.11-2.29	30
Band 8 — Panchromatic	0.50-0.68	15
Band 9 — Cirrus	1.39-1.38	30
Band 10 — TIRS	10.6-11.19	100 (30)
Band 11 — TIRS	11.50-12.51	100 (30)

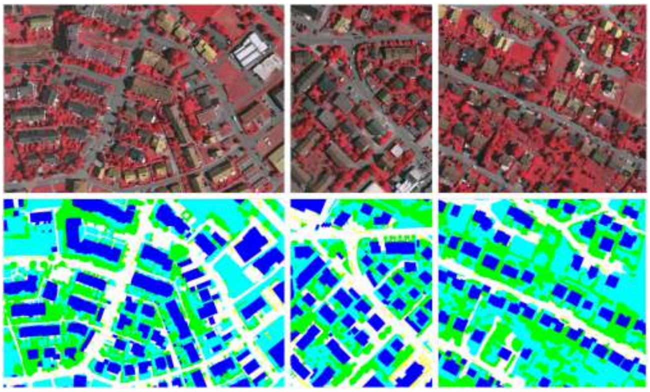


Fig. 8. Sample RGB images from Vaihingen dataset (top) and their ground truth (bottom).

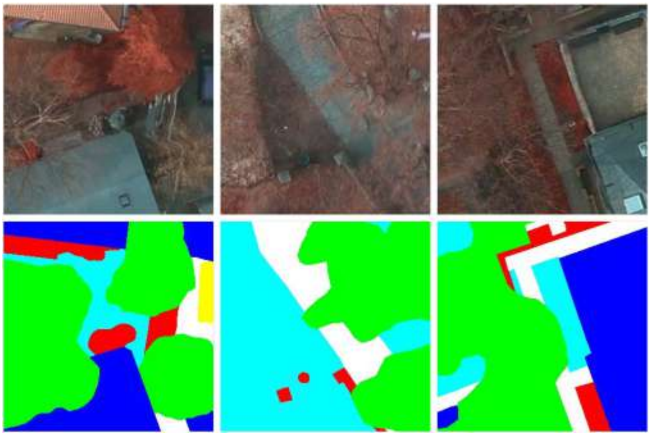


Fig. 9. Sample RGB images from Potsdam dataset (top) and their ground truth (bottom).

the TOP and the DSMs is 5 cm, differing from the 9 cm in Vaihingen. The dataset comprises 38 patches of 6000 × 6000 pixels, is available in TIFF format, and offers three channel compositions: NIR-RG, RGB, and RGB-NIR, where each channel has an 8-bit

TABLE VIII
OPERATIONAL FEATURES OF SENTINEL MISSIONS

Mission	**Launch date**	**End of operations**	**Repeat cycle**	**Instruments**
Gaofen-1	26 April 2013	31 December 2023	41 days	Panchromatic and Multispectral Camera (PMC) + Wide Field Imager (WFI)
Gaofen-2	19 August 2014	31 December 2023	69 days	PMC

TABLE IX
DESCRIPTION OF GAOFEN BANDS

Spectral bands	**Wavelength** (μm)	**Resolution** (m)
Gaofen-1		
Panchromatic	0.45-0.90	2
Band 1—Blue	0.45-0.52	8
Band 2—Green	0.52-0.59	8
Band 3—Red	0.63-0.69	8
Band 4—NIR	0.77-0.89	8
Gaofen-2		
Panchromatic	0.45-0.90	0.8
Band 1—Blue	0.45-0.52	3.2
Band 2—Green	0.52-0.59	3.2
Band 3—Red	0.63-0.69	3.2
Band 4—NIR	0.77-0.89	3.2

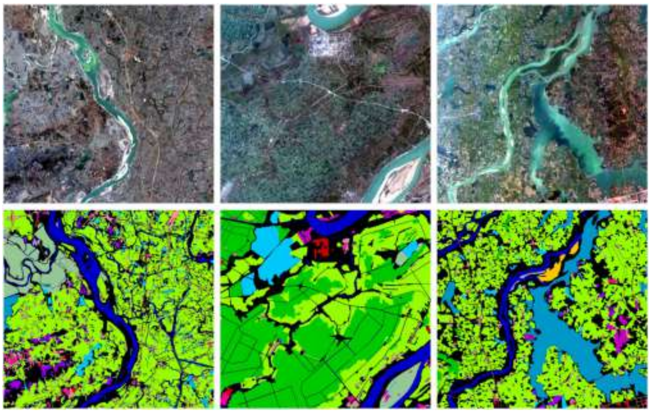


Fig. 10. Sample RGB images from GID dataset (top) and their ground truth (bottom).

spectral resolution. In addition, the images have been annotated at the pixel level and contain the same categories as Vaihingen.

3) Gaofen Image Dataset: The Gaofen Image Dataset (GID) [115] ranks as the third most frequently utilized benchmark in the studies comprising our review, which, like Potsdam and Vaihingen, is also focused on land cover. The images within GID were sourced through the GF-2 satellite service, covering more than 60 different cities across China.

GID is divided into two parts: a large-scale classification set and a fine LCC set. The large-scale classification set contains 150 images, while the fine classification set comprises 30 000 image patches, spanning a broad spectrum of geographical areas and landscapes, as shown in Fig. 10. The images feature four MS bands: RGB and NIR. The image size in the large-scale subset is 6800 × 7200, whereas the fine subset is multiscale, with sizes of 56 × 56, 112 × 112, and 224 × 224. The large-scale subset includes pixel-level annotations for five categories: built-up, farmland, forest, meadow, and water; the fine classification subset, meanwhile, provides annotations for 15 categories: paddy field, irrigated land, dry cropland, garden land, arbor forest, shrub land, natural meadow, artificial meadow, industrial land, urban residential, rural residential, traffic land, river, lake, and pond. The dataset is available for open access via the project's webpage.

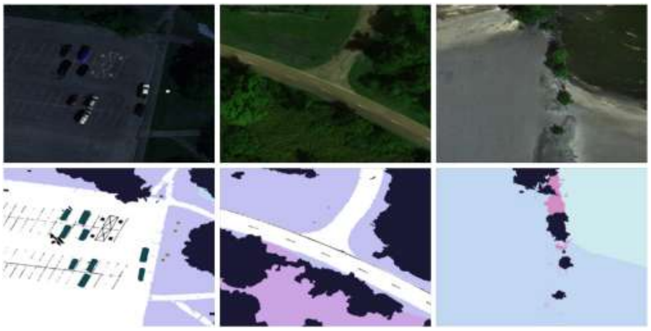


Fig. 11. Sample images from RIT-18 dataset (top) and their ground truths (bottom).

4) RIT-18: Another dataset identified in our review is RIT-18 [116], which is less frequently used compared to the datasets mentioned earlier. Examples of this dataset are shown in Fig. 11. This dataset was collected at Hamlin Beach State Park, located in the state of New York, USA, using a Tetracam Micro-MCA6 mounted on a DJI-S1000 octocopter. It offers six spectral bands, including RGB and three NIR bands covering a spectral range from 490 to 900 nm. In addition, it features a GSD of 4.7 cm. RIT-18 is annotated with 18 classes: road markings, tree, building, vehicle, person, lifeguard chair, picnic table, black panel, white panel, orange pad, buoy, rocks, low vegetation, grass/lawn, sand/beach, water (lake), water (pond), and asphalt. Notably, the dataset exhibits a severely unbalanced class distribution, posing a significant challenge for LCC methods. It comprises three images: one for training (9393 × 5642), one for validation (8833 × 6918), and one for testing (12 446 × 7654). The dataset is provided in the form of NumPy arrays and is available for free download and unrestricted use at GitHub.[18] The dataset size is approximately 3 GB.

[18] [Online]. Available: https://github.com/rmkemker/RIT-18

TABLE X
FEATURES OF OTHER SATELLITE IMAGING SERVICES

Service	Number of bands	MS bands	MS (m)	Resolution	Access type
Pléiades	1 PAN, 4 MS	RGB-NIR	2		Restricted
Ziyuan-3	4 MS	RGB-NIR	6		Commercial
SPOT-6/7	1 PAN, 4 MS	RGB-NIR	8		Commercial

TABLE XI
SUMMARY OF ALL MS BENCHMARK DATASETS FOR LAND COVER IDENTIFIED IN THIS REVIEW

Dataset	Number of classes	Spectrum/Bands	Number of images	Image size
ISPRS Vaihingen[a]	6	RG-NIR	38	≈2494×2064
ISPRS Potsdam[b]	6	RGB-NIR	33	6000×6000
GID[c]	15	RGB-NIR	150	6800×7200

RIT-18[d]	18	RGB-NIR	3	9393×5642;
				8833×6918;
				12 446×7654

SEN12MS[e]	33	Sentinel-2	541 986	256×256
Urban Semantic 3D[f]	5	WorldView-3	69	1024×1024

[a] [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx
[b] [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx
[c] [Online]. Available: https://captain-whu.github.io/GID15/
[d] [Online]. Available: https://github.com/rmkemker/RIT-18
[e] [Online]. Available: https://github.com/schmitt-muc/SEN12MS
[f] [Online]. Available: https://ieee-dataport.org/open-access/data-fusion-contest-2019-dfc2019

5) SEN12MS: The SEN12MS dataset [117] comprises 180 662 patch triplets (541 986 total images) incorporating Sentinel-1 dual-pol SAR, Sentinel-2 MS, and MODIS land cover data. These patches are sampled from randomly selected regions of interest across four meteorological seasons: winter (1 December 2016 to 28 February 2017), spring (1 March 2017 to 30 May 2017), summer (1 June 2017 to 31 August 2017), and fall (1 September 2017 to 30 November 2017). The dataset offers patches, with each one standardized to a size of 256 × 256 pixels in TIFF format. Moreover, it features a GSD of 10 m and encompasses 13 spectral bands. This dataset was compiled from diverse scenes distributed globally. For LCC purposes, it typically adheres to a scheme of the International Geosphere Biosphere Programme, consisting of ten classes: forest, shrubland, savanna, grassland, wetland, cropland, urban, snow, barren, and water. SEN12MS is freely available for download from a GitHub repository.[19] Fig. 12 shows some examples from this dataset.

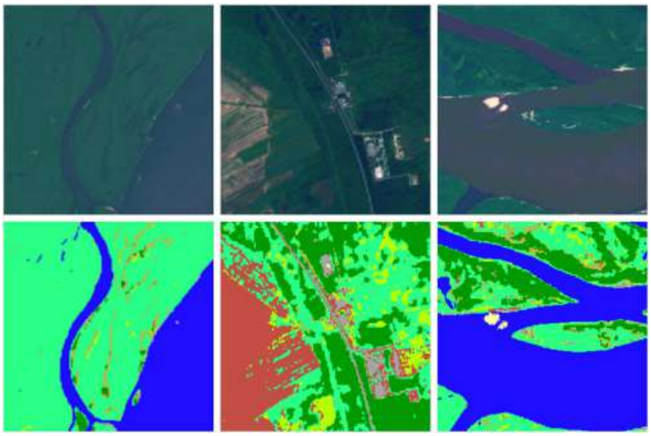


Fig. 12. Sample images from SEN12MS dataset (top) and their ground truths (bottom).

6) Urban Semantic 3D: The Urban Semantic 3D (US3D) dataset [118] encompasses satellite images, airborne LiDAR data, and semantic labels across approximately 100 km^2 in Jacksonville, FL, and Omaha, NE, USA. It features WorldView-3 panchromatic images along with eight-band visible and NIR (VNIR) images. The GSD is about 35 cm for panchromatic images and 1.3 m for VNIR images. The dataset comprises 69 images with a resolution of 1024 × 1024 pixels, each accompanied by corresponding semantic labels categorizing ground, vegetation, building, water, and elevated road classes. The dataset is available in TIFF format and can be accessed via IEEE DataPort.[20] Examples from the dataset are shown in Fig. 13.

[19] [Online]. Available: https://github.com/schmitt-muc/SEN12MS

[20] [Online]. Available: https://ieee-dataport.org/open-access/data-fusion-contest-2019-dfc2019

TABLE XII
SUMMARY OF UAV-CAPTURED DATASETS UTILIZED IN REVIEWED ARTICLES

Device/Sensor	Spectrum/Bands	Mounting platform	Number of classes	Ref.
Integrated camera with 1" CMOS 20M sensor	RGB-NIR	DJI P4 RTK Drone	6	[119]
MicaSense RedEdge-M WIRIS Pro SC	RGB-NIR-RE Thermal	DJI M600 Pro Drone	5	[31]
SlantRange MS sensor	RGB-NIR	Geo-X8000 Octocopter Drone	5	[45]

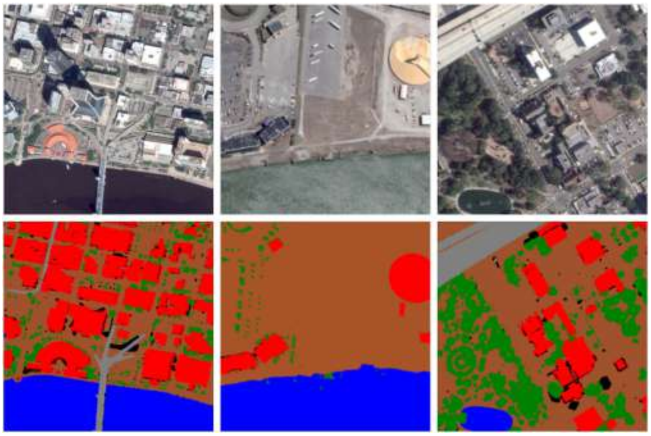


Fig. 13. Sample images from US3D dataset (top) and their ground truths (bottom).

C. Unmanned-Aerial-Vehicle-Captured Datasets

Unmanned aerial vehicle (UAV)-captured datasets are becoming an integral part of research methodologies. Utilizing drones offers a flexible and efficient means of data collection, allowing researchers to access difficult terrains and capture data at fine spatial resolutions. In this review, it was noted that a few studies successfully employed UAV-captured datasets, taking advantage of their ability to provide detailed and current data. For instance, DJI drones, equipped with MS cameras, have been highlighted as a popular choice among researchers for their versatility and ease of integration. Further details on these observations can be found in Table XII.

Nevertheless, despite their advantages, UAVs may not always be suitable for all types of data acquisition, especially in the context of large-scale land cover studies. Traditional land cover datasets often require a broad aerial view to capture extensive areas, a task that might be challenging for UAVs due to their limited range and payload capacity. In addition, UAVs can face restrictions in terms of flight duration and weather conditions, which may affect their ability to gather data over large continuous landscapes.

Regardless, UAVs are increasingly becoming a valuable tool across various application fields, including LCC. Their capacity to provide rapid detailed insights into specific locations makes them indispensable for many modern research and practical applications, complementing traditional methods and providing critical data that would otherwise be difficult to obtain.

VI. PREPROCESSING METHODS

Although no new or particular preprocessing methods were found in our review, it is important to highlight some common techniques that remain essential. These preprocessing steps, despite being well known, deserve description due to their significant impact on the quality and utility of the data used for MSSS in LCC. This section will address these procedures, emphasizing their importance and best practices in their application.

A. Image Correction and Enhancement

The first group of preprocessing techniques identified in this review pertains to satellite-obtained images. These techniques are particularly relevant in studies that utilize satellite imagery for specific problems. This differs from the use of benchmark datasets, as many of these are already prepared and relatively ready for analysis, requiring little or no additional preprocessing. In contrast, satellite images frequently require adjustments to correct certain aspects before they can be processed and analyzed, given their nature.

The first preprocessing technique to discuss is atmospheric correction. This is an important procedure in the preprocessing of remote sensing images, employed to mitigate the effects of the atmosphere that distort data captured from space [120]. This distortion occurs because the earth's atmosphere, composed of various gases and particulate matter, can absorb and scatter light reflected from the earth's surface before it reaches satellite sensors [121], [122], distorting the true appearance of the surface in the images. These atmospheric phenomena alter reflectance and can introduce artifacts such as haze or solar glare. The goal of atmospheric correction is to determine the earth's surface's actual reflectance from the apparent reflectance measured by the sensor [123]. To achieve this, atmospheric correction methods adjust the image's reflectance values to more accurately reflect the surface's actual characteristics. This way, the visual quality of the image is improved, and the intrusive component of the atmosphere is eliminated. An example of the effect of this technique can be observed in Fig. 14.

The next preprocessing technique for remote sensing images is radiometric correction, also known as radiometric calibration. This process ensures that the pixel values in images captured from space accurately represent the true reflectance of the earth's surface [124]. During image capture, various factors such as variations in sensor response, calibration errors of the instrument [125], or electronic interferences can distort the

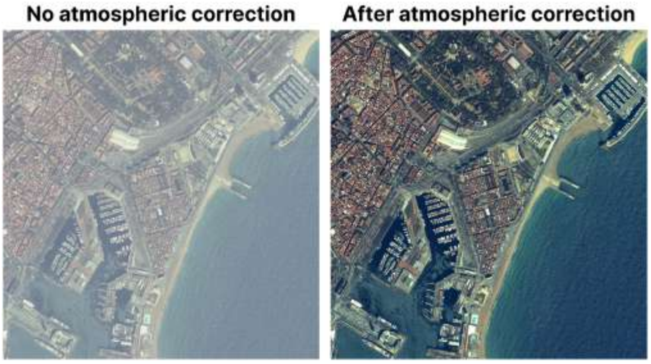


Fig. 14. Comparative example of the effect of atmospheric correction on a satellite image.

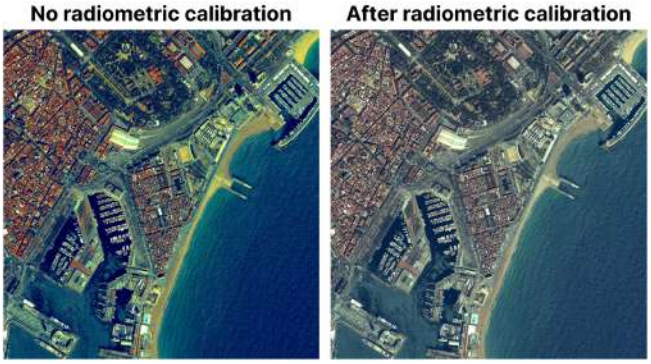


Fig. 15. Comparative example of the effect of radiometric correction on a satellite image.

data [126]. These errors may manifest as incorrect pixel values, which, for example, could cause images to appear overexposed or underexposed [127], thus affecting the accuracy of spectral measurements. Radiometric correction adjusts these values to correct such distortions, aligning the data to reflect the actual observed conditions. In other words, it achieves uniform intensity values [124], typically accomplished by balancing histograms or correcting imperfections present in the pixels [128]. Thus, this process improves the visual quality and reliability of satellite images, enabling a more precise interpretation of the terrain. Fig. 15 shows an example of the application of this technique.

Moving forward, another important technique is orthorectification. Occasionally, due to terrain topography, camera angle, or the movement of the satellite or vehicle capturing the image, distortions can occur in the alignment of images and their pixels [129]. Orthorectification eliminates the geometric and scale errors in images [130], ensuring that they are correctly located geographically. This is done by transforming the image's central projection system into an orthogonal projection [131], aiming to remove displacements caused by sensor movement and terrain relief. As a result, each pixel is correctly aligned with the actual geographic coordinates on the ground [132]. The outcome of this process is an image with cartographic precision and consistent planimetric scale. This is essential for the images to be used for precise mapping, measuring actual distances, and other geospatial applications where spatial location accuracy is critical.

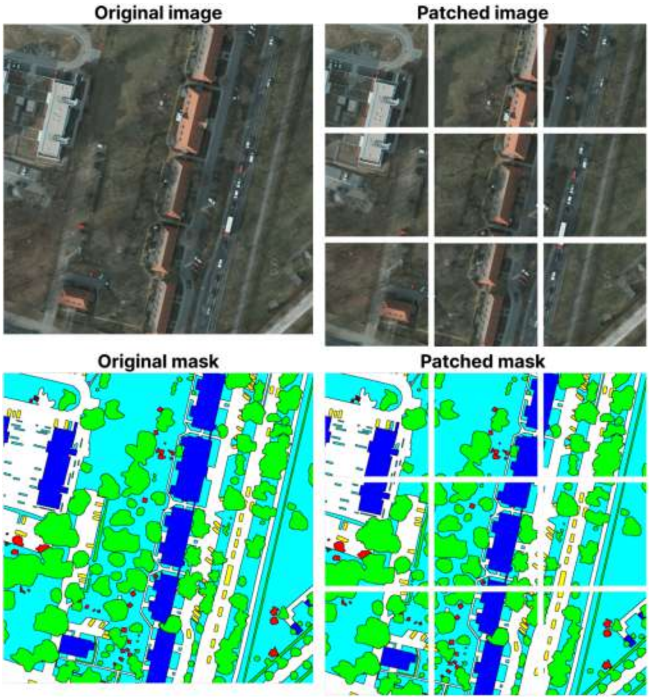


Fig. 16. Example of patching a large-scale image and its segmentation mask/ground truth.

The last technique is normalization. This process aims to standardize image data to ensure consistent conditions across captures [133]. Typically, image capture involves various devices and sensors, each with unique characteristics that can result in differing image features. Normalization is crucial for making disparate data comparable. It adjusts the images so that pixel values are consistent across different lighting conditions, capture angles, and sensors, eliminating variations that might be caused by external factors rather than actual changes on the observed surface [134]. This is especially vital in long-term tracking studies, such as change detection, as it enables more accurate and reliable analysis of changes over time and space.

B. *Large-Scale Image Handling*

In LCC tasks, images often come in large sizes, making it challenging to use them directly with standard methods or models. Processing these large-scale images in their entirety can be computationally expensive and inefficient [135]. Therefore, specific techniques are applied to handle these large images effectively, ensuring that the analysis remains feasible and accurate. This section focuses on the strategies employed to manage and process large-scale images in LCC, identified in this review.

The typical approach for handling large-scale images is to divide these into smaller patches [136], as shown in Fig. 16. This technique, known as patching, allows for more efficient and feasible management of large images for LCC methods and models [137], [138]. By splitting the image into patches, the computational load is reduced, as each smaller patch can

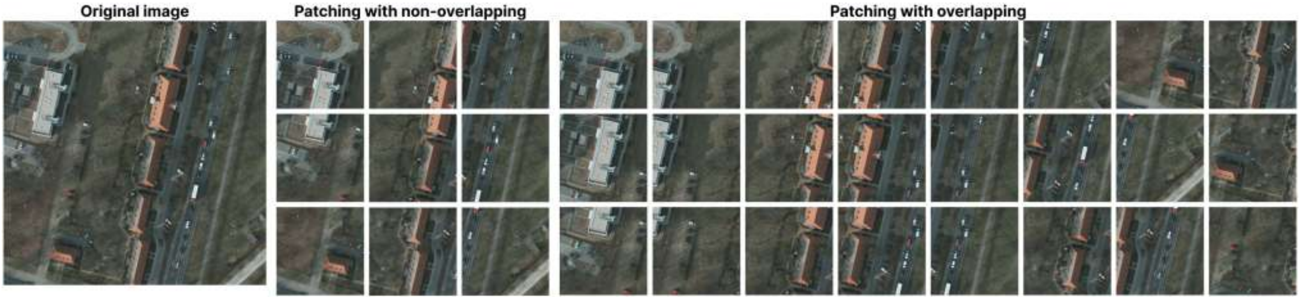


Fig. 17. Example of patching a large-scale image with nonoverlapping and overlapping.

TABLE XIII
DIVISION OF THE TRAINING AND EVALUATION SETS OF THE POTSDAM AND VAIHINGEN DATASETS ACCORDING TO THEIR IMAGE ID

Split	Potsdam	Vaihingen
Training	2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_10, 7_11, 7_12	1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, 37
Evaluation	2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 5_13, 5_14, 6_13, 7_13	2, 6, 12, 16, 22, 27, 31, 35

be processed individually. This not only facilitates parallel processing but also enables models to focus on specific areas of the image, potentially improving segmentation and classification accuracy. Patching is also beneficial for addressing memory issues since large images can exceed the memory capacity of processing systems [138]. Working with smaller patches ensures that operations can be performed without exhausting available resources.

Regarding the size of the patches, there is no fixed standard; it depends on the desired level of detail and the specific requirements of the model or technique being used. This variation can occur whether a custom model is implemented with a specific input size or preexisting methods are employed, as these typically necessitate specific input requirements. Generally, the ideal patch size should be small enough to facilitate efficient processing but large enough to capture relevant landscape features. Some common sizes used, according to what was found in this review, are 512 × 512 and 256 × 256. Furthermore, some datasets, such as Potsdam and Vaihingen, designate specific images for training and others for evaluation, as shown in Table XIII, which are then patched accordingly.

Moreover, large-scale images can be patched with nonoverlapping or overlapping, as shown in Fig. 17. Nonoverlapping patches are created by dividing the image into segments using a stride that matches the chosen patch size. This method is straightforward and reduces computational load, but it can result in the loss of important semantic information at the edges of the patches. Alternatively, patches can be created with overlapping, where adjacent patches share some pixels. This is achieved by adjusting the stride to a size smaller than the chosen patch size. This approach helps to reduce the loss of semantic information during the patching process [139], ensuring that important features are captured in multiple patches. Overlapping patches can maintain the continuity and context of the data, which is crucial for accurate segmentation and classification.

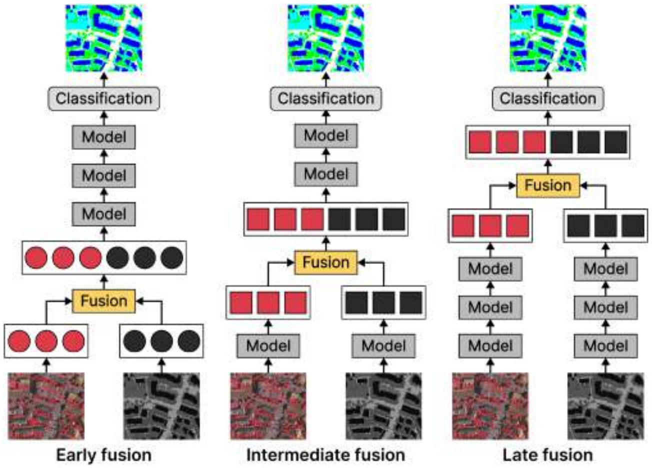


Fig. 18. Graphical representation of image fusion methods.

Finally, it is important to note that there are specific considerations when patching MS images. This is due to the fact that common tools and image processing software may not be equipped to handle images with more than three channels. Moreover, as discussed in the previous section, these types of images usually come in formats such as TIFF rather than the more common JPEG or PNG formats. Therefore, it is crucial to ensure that when patching an MS image, the process is done correctly and that the patches maintain all the spectral channels. Failure to do so can result in the loss of critical information and may adversely affect the analysis and classification outcomes.

C. Channel Fusion

Another common preprocessing step in LCC using MSI is channel fusion. Channel fusion involves combining multiple spectral bands to create a unified representation of the data [140], which can be useful for analysis. There are different approaches to channel fusion. As shown in Fig. 18, channels can be fused early in the preprocessing stage, before any further main processing is done, processed separately and fused later, or even fused at an intermediate step [141], [142]. Here, to be consistent with the

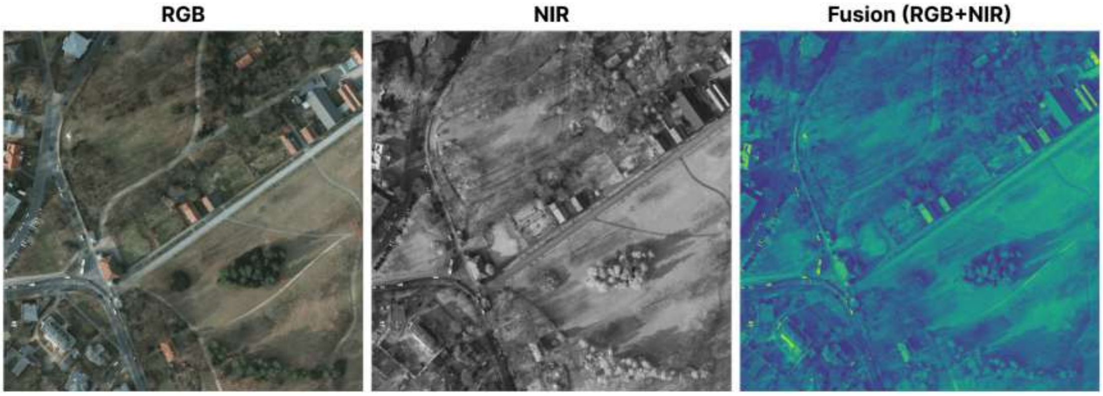


Fig. 19. Graphical visualization of the fusion of RGB channels with NIR using averaging on an image from the Potsdam dataset.

essence of preprocessing, we will refer to channel fusion as any fusion step that can be performed before the main processing steps.

In the area of LCC, channel fusion can involve combining visible bands with nonvisible bands to generate a single representation. For example, visible bands can be fused with NIR or thermal bands to provide a more comprehensive view of the landscape [143], [144], as shown in Fig. 19. Specific bands can be fused to highlight particular features of interest, such as vegetation or water bodies. In addition, other sources of spectral information, such as DSMs or spectral indices like NDVI, can be included in the fusion process [145], [146]. By integrating DSMs, the fused data can incorporate elevation information alongside spectral data, which enhances the overall representation. Similarly, including spectral indices such as NDVI allows the fusion process to incorporate metrics of vegetation health [145], adding further detail and context. This comprehensive integration of multiple data modalities can help improve the interpretation of the terrain and achieve better classification results.

There are various techniques available for channel fusion, each presenting unique approaches and advantages. A widely known method involves concatenation, where multiple channels are merged by stacking them [147]. This approach preserves the original values of each channel, creating a dataset that incorporates all channels as distinct layers. It is straightforward to implement and ensures that all the initial information from each channel is maintained. Weighted sum is a variant, where each channel is assigned a specific weight based on its importance or relevance [148]. The weighted values of corresponding pixels are then summed to produce a single fused channel. Another common approach is averaging [149], where values from corresponding pixels across multiple channels are averaged to generate a unified channel. This method decreases data volume and noise while emphasizing shared features among channels. Weighted averaging is a variant, assigning weights to channels based on their significance.

Dimensionality reduction is another approach that can be used for channel fusion [150]. This involves transforming the original channels into a smaller set of components that retain most of the important information [151]. For example, the well-known principal component analysis (PCA) transforms the original channels into a set of linearly uncorrelated components while retaining as much significant information as possible [152]. This method can facilitate the handling of high-resolution images, improve efficiency, and highlight the most significant patterns in the data. There are also more advanced methods such as the wavelet transform, which decomposes channels into different frequency components [153]. These components are then fused at various scales, capturing both spatial and frequency information. This approach proves particularly advantageous for conducting comprehensive multiscale analyses, enabling a thorough exploration of both spatial features and frequency characteristics within the data.

VII. Evaluation Metrics

In the realm of MSSS for LCC, the evaluation of models against ground truth data is necessary for assessing their accuracy and effectiveness. This process involves comparing the segmented outputs generated by the models with ground-truth segmentation masks to determine how closely they match. Given the importance of this evaluation, it becomes essential to employ metrics that can accurately quantify the performance of semantic segmentation algorithms.

In this section, we delve into the evaluation metrics that, according to our review, are most commonly used in our field of study. While many of these metrics are universally applied across various segmentation tasks, their importance in validating the results against ground-truth data remains paramount. In the following sections, we will detail each metric, providing explanations and interpretations.

A. *Overall Accuracy*

Overall accuracy (OA) is a straightforward metric that quantifies the proportion of correctly classified pixels over the total number of pixels in the image [23], [154]. It serves as a primary measure for evaluating the overall effectiveness of semantic segmentation models, especially when assessing their capability to accurately distinguish between different classes across the entire image [155]. A higher OA indicates a greater number of pixels correctly classified, reflecting the model's general performance in semantic segmentation tasks. Mathematically, OA is calculated as the sum of correctly predicted pixels for all classes divided by the total number of pixels in the image, as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

where TP represents the true positives, TN the true negatives, FP the false positives, and FN the false negatives.

In essence, a model with high OA has a greater capability to segment the image accurately, aligning closely with the ground truth. However, it is important to note that while high OA can indicate effective segmentation, this metric may not fully capture the model's performance in handling class imbalance, where certain classes are underrepresented in the dataset. Therefore, OA should be considered alongside other metrics to provide a more comprehensive evaluation of the model's capabilities in semantic segmentation.

B. Kappa Coefficient

The kappa coefficient, also known as kappa score, is a statistic that measures interrater reliability for qualitative (categorical) items [156]. It assess the agreement between the predicted labels by a model and the ground truth, correcting for the agreement that could occur by chance [157]. This metric provides a more nuanced evaluation of model accuracy than simple percent agreement, as it considers both the agreement on positive instances and the potential for random agreement on negative instances. Mathematically, the kappa coefficient is calculated as

$$\mathrm{kappa} = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

where P_o is the observed agreement among raters, and P_e is the hypothetical probability of chance agreement, calculated using the frequencies of each category [157], [158].

High values of kappa show high agreement between what the model predicts and the ground truth, suggesting that the model is performing well. Conversely, low values of kappa would suggest that much of the agreement could, in fact, be due to chance and thus point to poor model performance. Values of kappa range from -1 (perfect disagreement) through 0 (agreement equivalent to chance) to 1 (perfect agreement) [157].

C. Mean Pixel Accuracy

Mean pixel accuracy (mPA) is an important metric for assessing the performance of semantic segmentation models across various classes, especially useful in datasets with class imbalance. This metric computes the accuracy individually for each class and then averages these values, ensuring a balanced evaluation that considers the performance of the model on each class independently [16], [159]. Mathematically, mPA is calculated as follows:

$$\mathrm{mPA} = \frac{1}{C}\sum_{i=1}^{C} \frac{p_{ii}}{\sum_{j=1}^{C} p_{ij}} \tag{3}$$

where p_{ii} is the number of pixel of class i predicted to belong to class j, and C is the number of categories.

A high mPA indicates that the model not only performs well across all classes on average but also demonstrates consistent accuracy and fairness in classifying different classes, regardless of their frequency. Conversely, a low mPA points to inconsistencies in the model's ability to accurately classify different classes.

D. Precision

Precision, often referred to as the positive predictive value, quantifies the accuracy of the positively predicted pixels or segments within a specific class [160]. It is a crucial metric for semantic segmentation models, especially when the cost of FP is high. Precision is particularly insightful in datasets with imbalanced classes [161], highlighting the model's ability to correctly identify relevant pixels amidst a majority of irrelevant ones. Mathematically, precision is defined as the ratio of TP to the sum of TP and FP, as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{4}$$

High precision indicates that the model is effective in segmenting pixels as belonging to the target class, with minimal misclassification of other-class pixels as target class. This metric is particularly valuable in scenarios where the objective is to minimize incorrect segmentations of a specific class, even if it means potentially missing some TP.

E. Recall

Recall, also known as sensitivity or true positive rate, measures the model's ability to correctly identify all relevant instances within a specific class [162]. This metric is paramount in situations where failing to detect an instance of the target class carries significant consequences. Defined as the ratio of TP to the sum of TP and FN, recall provides insight into the model's comprehensiveness in capturing the target class instances [23], as illustrated in the following equation:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{5}$$

A model with high recall efficiently detects the majority, if not all, true instances of a particular class, ensuring minimal misses. While high recall is desirable, it is often achieved at the expense of precision, as efforts to reduce missed detections can lead to an increase in FP. Therefore, recall should be evaluated in conjunction with precision to achieve a balance that suits the specific requirements of the application at hand, enabling a more nuanced understanding of the model's performance tradeoffs.

F. F1-Score

The F1-score is a harmonic mean of precision and recall, providing a single metric that balances both the model's ability to correctly identify positive instances and its ability to find all positive instances [158], [163]. This metric is particularly useful in scenarios where both FP and FN have significant implications, and neither precision nor recall can be prioritized over the other [164]. It offers a way to evaluate the model's overall performance in semantic segmentation tasks, especially when dealing with imbalanced datasets. Mathematically, the F1-score is calculated as follows:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

Achieving a high F1-score is indicative of a model's balanced performance across its segmentation tasks, managing to both accurately classify pixels as belonging to the target class and encompassing the majority of those instances.

1) Mean F1-Score: The mean F1-score (mF1) averages the F1-scores of each class, offering a balanced metric for multiclass segmentation tasks [154]. It effectively captures the model's performance across all classes, especially useful in handling class imbalances. Mathematically, the mF1 is computed as follows:

$$\text{mF1} = \frac{1}{C}\sum_{i=1}^{C}\left(2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}\right) \tag{7}$$

where C is the number of classes, and Precision_i and Recall_i measure the correct classification and completeness of the detection for each class, respectively. A high mF1 score indicates that the model performs well and consistently across different classes, achieving a balance between identifying relevant instances and minimizing oversegmentation. It is particularly valuable for assessing performance in datasets where every class is critical.

G. Intersection Over Union

Intersection over union (IoU), also known as the Jaccard index [163], is a fundamental metric for evaluating the accuracy of semantic segmentation models. It measures the overlap between the predicted segmentation and the ground truth, providing a clear indication of the model's precision in delineating target areas [156]. The IoU is especially valuable for its ability to balance the impact of TP predictions against FP and FN predictions, making it a comprehensive and robust metric for segmentation tasks. Mathematically, this is represented as follows:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{8}$$

where $|A \cap B|$ represents the intersection, which is the number of pixels correctly identified as belonging to the target class by both the model's prediction and the ground truth. On the other hand, $|A \cup B|$ denotes the union or the total number of unique pixels that have been classified as the target class in either the prediction or the ground truth.

The IoU ranges from 0 to 1, where 0 indicates no overlap and 1 represents perfect segmentation alignment with the ground truth. Thus, models with scores closer to 1 are better at identifying and classifying all relevant pixels, ensuring that the segmented output closely matches the ground truth.

1) Mean Intersection Over Union: The mean intersection over union (mIoU) extends the concept of IoU by averaging the IoU scores across all classes, offering a comprehensive metric that evaluates the model's segmentation performance over the entire dataset [156]. This metric is especially beneficial in scenarios with multiple classes, as it provides a holistic view of the model's effectiveness in segmenting each class relative to the ground truth. The mIoU is calculated by first computing the IoU for each class, then averaging these scores [23], [156]. The formula is given as follows:

$$\text{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\text{IoU}_i = \frac{1}{C}\left(\frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}\right) \tag{9}$$

where IoU_i is the IoU score for class i, and C is the total number of classes.

mIoU scores also range from 0 to 1, with values closer to 1 indicating superior segmentation performance across all classes. A high mIoU signifies not only that the model accurately identifies the pixels belonging to each class but also that it effectively minimizes misclassifications across the board. Conversely, a lower mIoU score suggests that the model's segmentation results are inconsistent or generally inaccurate across different classes.

VIII. Methods and Paradigms

In this section, we describe the various MSSS methods and paradigms for LCC that have been identified in the reviewed papers. Our goal is to analyze the techniques and approaches found in recent works that leverage MSI data to address LCC challenges. This aims to provide a clear understanding of the current methodologies and their applications in the field, illustrating the advancements and innovations that characterize contemporary practices.

A. Machine Learning

Despite the growing trend toward the adoption of DL techniques in various CV tasks, including MSSS for LCC, our review has identified a subset of research that continues to employ traditional ML approaches. Although these methods do not represent the majority of recent work, their presence underscores a continued relevance in specific LCC contexts. The persistence of these techniques underlines a methodological diversity that enriches the research landscape, offering robust and efficient alternatives for LCC. Within this group of reviewed works, approaches based specifically on OBIA and PBIA with a primary use of ML classifiers were found.

A prime example is identified in [165], where a case of LCC in Johannesburg is addressed. The authors adopt the OBIA approach, starting with a multiresolution segmentation (MRS) algorithm that facilitates the grouping of pixels into coherent objects at multiple scales [166], allowing for the precise detection of both fine and coarse features. This initial segmentation serves as an essential preparatory step, outlining significant analytical units without yet assigning them a semantic identity. Classifiers are then used to assign each object a class or category based on its characteristics. In this case, the effectiveness of two classifiers is evaluated: SVM and RF, applied to MSI images with eight spectral bands. In [167], another land cover mapping case study is conducted, this time in Eyüpsultan, Türkiye, using SAR imagery and 12-band MSI imagery. For mapping, an OBIA approach is followed, where initially an MRS algorithm is applied to divide the images into homogeneous areas. Subsequently, three classifiers, i.e., SVM, RF, and K-nearest neighbors (KNN), are employed for semantic identification. In addition, NDVI is used as supplementary information to enhance the identification of vegetated areas. In the same way, in [168], another study is conducted for LCC in the suburb of Malad Creek, Mumbai, India, using MSI images consisting of eight bands. An OBIA approach is used, where initially an MRS algorithm is applied to delineate the zones. Prior to classification, PCA is applied to reduce the data's dimensionality. Finally, KNN and RF are evaluated and compared for semantic classification.

Moving forward, the study conducted in [45] focuses on LCC in the southwestern coastal area of Bangladesh, particularly in the city of Barishal, Bangladesh. Images used are four bands: RGB and NIR. In terms of methodology, two approaches are evaluated: PBIA and OBIA. For the first approach, involving direct application of the classifier, K-means and maximum likelihood classifier are utilized. For the OBIA approach, the MRS algorithm combined with KNN is employed. In addition, in the second approach, NDVI, NDWI, and MSAVI2 indices are utilized. Similarly, in [169], an evaluation study of various ML approaches for mapping areas in northern Iran is conducted. For this, the authors used RG-NIR images. In addition, a PBIA approach is applied, in which SVM and RF classifiers were directly applied to produce the segmentation maps. Continuing with approaches based on pixel analysis, Zhao et al. [170] delve into LCC focusing on the urban area of Mardan, Pakistan. Their aim is to evaluate the efficacy of ML tree-based methods in producing precise classification maps, using MSI consisting of RGB, NIR, and SWIR bands. In addition, they incorporate data from the NDVI, MNDWI, and NDBI indices. In their exploration, the authors experiment with three ML models: classification and regression trees (CART), SVM, and RF. Employing a PBIA approach, these models were directly applied to the data to obtain the segmentation maps.

In this group of studies, a clear trend emerges: ML-based methods are often used for case studies in specific areas. This suggests that these methods are primarily employed to achieve quick results with minimal resource expenditure, allowing for rapid insights into the data. These approaches demonstrate the practical utility of ML techniques in LCC tasks, providing efficient and effective solutions. In addition, classifiers such as SVM and RF are frequently used, highlighting their effectiveness in handling MSI imagery without the need for extensive training datasets.

B. *Deep Learning*

This section analyzes the DL-based methodologies found in the reviewed literature. We have subdivided this into the main methods, architectures, and paradigms identified as recurrent. First, we examine the studies that apply existing methods directly to LCC tasks without significant modifications. Then, we analyze the approaches that implement minor modifications to preexisting methods. Following this, we transition to contributions that rely on attention mechanisms (AMs). Subsequently, we discuss the works that utilize transformers, adding another layer of complexity beyond AMs. After that, we explore studies that leverage multiscale information. Next, we review approaches that use edge information to enhance segmentation. Subsequently, we analyze generative method-based approaches. Finally, we examine approaches that process MS bands independently.

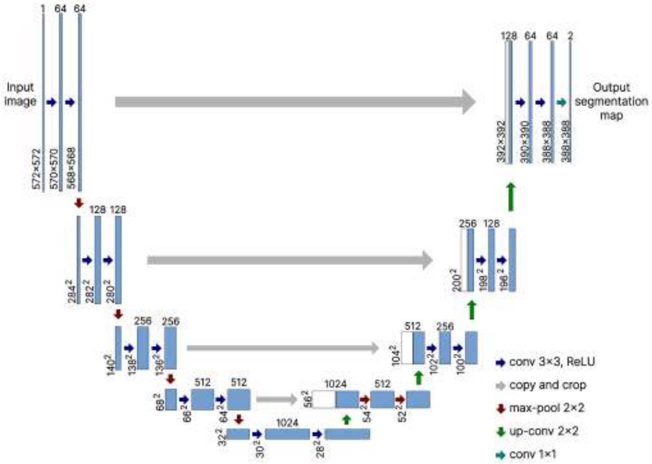


Fig. 20. UNet architecture.

Importantly, this subdivision is the result of the analysis of the extracted works using the criteria mentioned in Section III. In addition, we focus on the prominent methods and paradigms that have demonstrated significant advancements and applicability in the field of LCC using MSI. This approach ensures that the review captures the most relevant and impactful techniques currently being utilized, according to the extracted literature.

1) Off-the-Shelf Approaches: This segment of the review highlights research that leverages established DL models applied directly to address LCC using MSI. Unlike their counterparts in ML, these studies harness the intrinsic power of DL architectures, exploiting their sophisticated feature extraction and pattern recognition capabilities without the need for customization. Although few in number, these studies have emerged among the works reviewed, reflecting a reliance on the robustness and generalizability of preexisting models. These approaches showcase their immediate applicability and effectiveness in capturing the nuanced distinctions within diverse land cover types.

Within this group of reviewed works, UNet has emerged as the most commonly used architecture. UNet is a convolutional architecture originally designed for semantic segmentation tasks in medical images [82]. It is characterized by its U-shaped structure, as shown in Fig. 20, composed of a contracting path (encoder) to capture context and an expanding path (decoder) to allow for input reconstruction along with precise object localization [171]. This architecture facilitates network training with a limited number of images by preserving context and location information through skip connections between corresponding layers of the encoder and decoder, significantly improving segmentation accuracy.

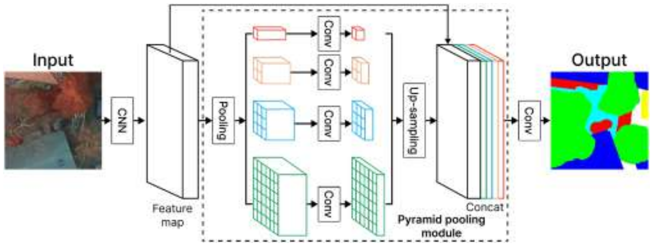


Fig. 21. PSPNet architecture.

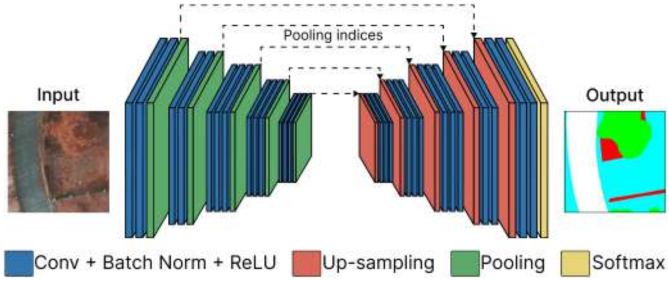


Fig. 22. SegNet architecture.

To begin with, Zheng and Chen [172], [173] explore the application of UNet using four-channel RGB-NIR images for land cover mapping in different regions of China. Li et al. [5] study the capability of UNet specifically in the Pidu area of China for natural resource analysis. The study includes a comparison of the impact of adding the NIR band in the segmentation process, in contrast to a classic RGB configuration. Giang et al. [119] conduct a land cover study in mining-affected areas in Daknong Province, Vietnam, using RGB-NIR images and UNet. In addition, they evaluated the use of different optimizers in the architecture.

Some works also utilize other well-known architectures in the world of CV for direct application in LCC. Chaurasia et al. [174], for example, map areas in the cities of Jacksonville and Omaha, USA, using data that contain panchromatic and eight-band VNIR images. In this case, a UNet and a pyramid scene parsing network (PSPNet) [175] are used to evaluate and compare their performances. PSPNet, shown in Fig. 21, is an architecture that consists of two main modules: first, a feature extractor, which can be another pretrained architecture, and a pyramid pooling module that gathers information at four different levels [176], which is then concatenated to generate the final prediction map.

In the same vein, Sathyanarayanan et al. [177] conduct LCC in areas of Mandya, India, using five-band images including RGB, NIR, and SWIR. The architecture selected for segmentation is SegNet [178], shown in Fig. 22, an encoder–decoder network composed of a series of convolutional layers, with its main innovation centered on the decoder. Specifically, it uses pooling indices generated during the max-pooling in the encoding stage to perform upsampling and generate the segmentation map. Sivagami et al. [179] also include SegNet in their methods for land cover mapping in urban areas of Germany using RG-NIR images. In addition, the study includes a 152-layer residual network (ResNet) [180], shown in Fig. 23, a CNN designed to address the vanishing gradient problem. ResNet includes shortcut connections that allow the gradient to skip some layers while flowing through the network [181], suffering less degradation, which consequently allows for the training of deeper networks. ResNet has several variants depending on the desired depth, with ResNet50, ResNet101, and ResNet152 being three of the most commonly used versions.

This group of studies highlights the use of certain architectures for direct application in LCC tasks using MSI. Most of these works are focused on case studies for specific areas and objectives and are not aimed at developing new methodologies. However, they provide a general overview of the applicability of the models and certain trends in their use, such as the recurrence of UNet and the prominence of PSPNet and ResNet. Despite being widely known architectures, they continue to be used, specifically for LCC, and are perfectly applicable for handling different bands and spectral information.

2) Off-the-Shelf Approaches With Slight Modifications: In this review, we also identified studies that apply small or minor modifications to existing architectures for LCC using MSI. Unlike more advanced techniques that may involve complex mechanisms or significant architectural overhauls, these approaches implement straightforward adjustments. These modifications include minor changes to the network architecture, such as adding or removing layers, and making slight alterations to some components of the architecture. Similar to the previous group of studies, this group is relatively small, but it highlights the potential impact of even minor adjustments on improving model performance to better handle the specific challenges of MSI.

First, Saxena et al. [182] introduce the Res-Seg-Net model, inspired by the architectures of ResNet and SegNet. The proposed model features an encoder–decoder structure and integrates the residual mapping of ResNet with the autoencoder approach of SegNet, without using fully connected layers. This aims to maintain high-resolution feature maps while reducing memory consumption and inference time for processing seven-band images.

Moving forward, in [134], a study is conducted on mapping mining sites in North America using UNet and RGB-NIR images. Slight changes are applied to the UNet, including reducing the number of filters in the convolutional layers and using elastic rectified linear unit (ReLU) instead of ReLU. In addition, NDVI and SAVI indices are included in the input as auxiliary information. Fan et al. [183] also use RGB-NIR images and UNet as a baseline for land cover mapping in areas of Hangzhou and Beijing, China. In this case, the modification involves replacing the encoder of the original UNet with a ResNet50 to develop Res-UNet. Furthermore, a second modification includes adding an inception module to the early layers of Res-UNet in search of additional performance improvement.

Another UNet-based approach for LCC using seven-band images, focusing on increasing the network's depth is proposed in [184]. However, instead of simply adding more layers, which can lead to overfitting, the authors decided to use 1 × 1 convolutional layers, allowing for an increase in depth without significantly increasing the number of parameters. In addition,

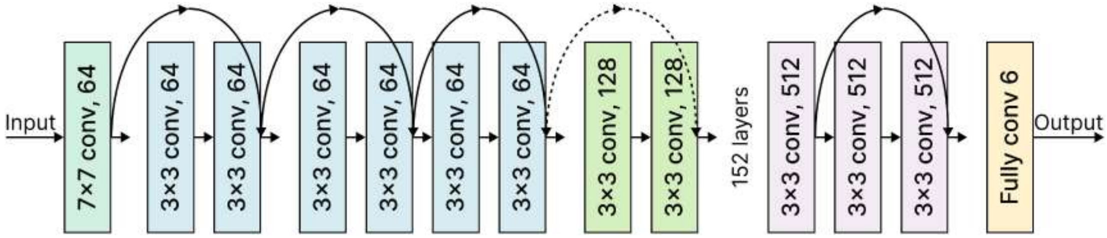


Fig. 23. ResNet152 architecture.

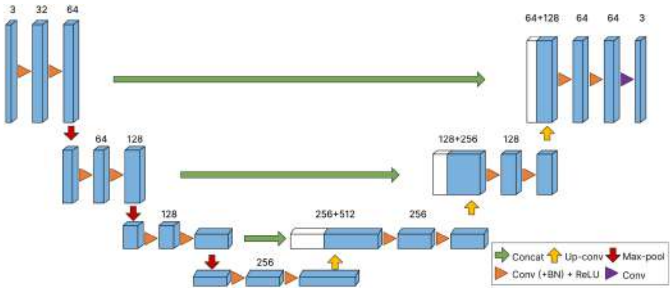


Fig. 24. UNet 3-D architecture.

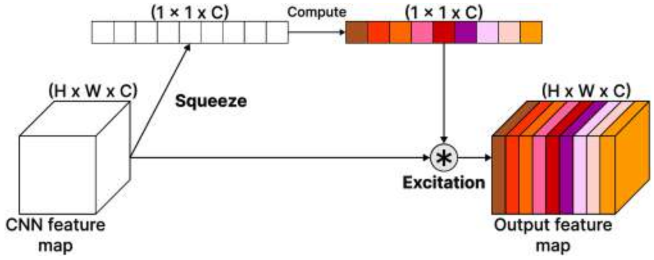


Fig. 25. Basic structure of an SE network.

they reduced the size of the kernels to 3 × 3, as they have the same effect as larger kernels. Pooling is also employed after each phase to reduce overfitting. To enhance the ability to propagate information from the early to the later layers, some deep supervised layers are added to supervise the network's training. Finally, padding layers and batch normalization layers are used to maintain network stability and improve the ability to capture data distribution.

Likewise, Aliyu et al. [185] aim to optimize the 3D-Unet architecture, shown in Fig. 24, which originally consists of 19 069 955 parameters, through strategies aimed at reducing bottlenecks and improving processing efficiency for LCC in seven-band MSI images. This optimization is achieved by increasing the number of channels before the max-pooling operation, allowing for a more refined analysis of the data. The methodology employs image tiles in voxel format of seven channels, resulting in precise segmentation in three dimensions, providing a broad context for effective learning of each voxel. In addition, batch normalization is incorporated for each ReLU activation, adjusting the mean and standard deviation during training and subsequently applying scale and bias adjustments based on global statistics during the testing phase. Furthermore, the model assigns null weights to unlabeled pixels, focusing learning on labeled data to ensure effectiveness and accuracy in segmentation.

In this group of studies, the recurrence of architectures, such as UNet and ResNet, is once again evident for the application of DL models with slight modifications. In addition, these works demonstrate how these architectures can be integrated to combine their best capabilities. By leveraging the strengths of each architecture without a significant increase in computational resources, these studies illustrate the potential of small adjustments to enhance model performance for LCC using MSI.

3) Attention Mechanisms: Within the techniques and paradigms identified in this review for LCC using MSI, AMs have been identified as significant tools. Initially, neural networks were designed to process inputs in a sequential or hierarchical manner, exemplified by approaches such as CNNs for image processing tasks and RNNs for sequential data. However, these traditional architectures faced limitations, particularly in capturing long-range dependencies and contextual relationships within data [186], [187]. This paved the way for the development of AMs, which have since revolutionized the landscape of DL. AMs are inspired by the cognitive processes of human perception that emphasize relevant signals while ignoring less useful ones. Fundamentally, the AM operates similarly, allowing neural networks to selectively focus on specific parts of the input data [188].

Initially, the AM was introduced with the aim of improving the performance of the encoder–decoder model used in machine translation [189]. This allows the decoder to effectively utilize the most relevant parts of the input sequence by employing a weighted combination of all encoded input vectors, assigning higher weights to the most relevant vectors. This proved highly effective for natural language processing (NLP) tasks, and shortly thereafter, they were adapted for CV, where AMs allow neural networks to selectively focus on specific parts of input images [10], thereby enhancing the ability to recognize and understand complex visual patterns. In MSSS, particularly for LCC, these AMs have been successfully applied. Their ability to capture complex contextual relationships allows for improved identification and differentiation among various land cover classes, resulting in more precise segmentation.

Several studies have demonstrated the effectiveness of AMs in LCC. For instance, Yang et al. [190] integrate an AM into each skip connection of a UNet network. This mechanism uses a squeeze-and-excitation (SE) [191] design (see Fig. 25) through pooling layers and fully connected layers to assign different weights to the channels of an RGB-NIR-NDVI input, fusing only the relevant information. Similarly, Nong et al. [192] introduce an AM in the skip connections of a UNet branch responsible for segmentation, applied to RGB-NIR-DSM and RG-NIR-DSM images. Another contribution [156] takes a different approach by integrating a channel attention module with discrete cosine transformation to extract global low-frequency features from RG-NIR and RGB-NIR images. This method allows for dynamic modeling of channel relationships across different frequencies, corresponding to various land cover types.

In some cases, SE-based AMs are used to bridge the gap between different feature levels. For example, Huang et al. [193] integrate an SE-based AM between the encoding stages of an encoder–decoder architecture to fuse higher level and lower level features, reducing the interlevel semantic gap from RG-NIR images. Similarly, Liang and Wang [194] employ a channel-level AM to control the fusion of higher level and lower level features and model the importance of the channels in four-band and seven-band MSI images. Likewise, in [195], AMs handle both low-level and high-level features using a dual-branch approach. Each branch integrates an SE-based patch-level attention module to enhance feature representation. In addition, an attention embedding module enriches low-level features by embedding local attention from high-level features. Finally, these enriched features are fused for improved segmentation from RGB-NIR-DSM and RG-NIR-DSM images.

Other approaches have focused on multiattention modules to enhance the extraction of spectral and spatial information from MSI. In [196], a multiattention module employs different pooling strategies and applies them to various groups of spectral bands (RGB, NIR, RE, and SWIR) to obtain rich features, enabling better LCC. Meanwhile, Chen et al. [197] employs two AMs: one for channel attention and another for spatial attention when processing RGB-NIR and RG-NIR images. The first focuses on interchannel relationships, while the second evaluates pixel connections across spatial locations. These two modules are connected in a cascade, with channel attention feeding into spatial attention, enhancing the perception of both spatial relationships and those between different MSI bands. A similar dual-attention approach is observed in [160], where features from a MobileNetV2 backbone are processed by a channel attention module followed by a spatial attention module. This sequence highlights significant features and minimizes less relevant ones from RGB-NIR images. Similarly, in [198], a framework with multiple AMs is proposed to improve MSI segmentation by considering not only spatial and channel information but also class-level information. It features two parallel branches: the class augmented attention module with class channel attention, and the region shuffle attention module. The class augmented attention calculates category-based correlations and generates weighted class representations, while the class channel attention recalibrates class-level information. The region shuffle attention captures regionwise global information with a shuffling operator, reducing feature redundancy and improving efficiency.

In [199], large kernel attention is incorporated within the encoder of a SegNet-based architecture to effectively capture long-range dependencies. The decoder of this architecture further enhances feature representation by employing a coordinate attention module that embeds positional information through pooling in horizontal and vertical directions, creating attention maps that result in better segmentation maps from RGB-NIR-DSM images. Building on kernel-based ideas, in [200], the dot product AM is rethought by generalizing as a kernel operation and using softplus instead of ReLU, forming a kernel attention module that reduces the complexity of the attention operation to a linear level. This module processes spatial features, while channel attention blocks model channel-level features from a ResNet50 backbone. Multiple attention blocks are used to process MSI inputs, achieving improved segmentation.

Overall, these studies illustrate that the use of AMs is prevalent in LCC using MSI. By selectively focusing on relevant parts of the input data and capturing complex contextual relationships, AMs provide a powerful means to improve the accuracy and efficiency of MSI segmentation approaches. Moreover, AMs are versatile, as they can be employed for various aspects, such as channel level or spatial level, to achieve better feature representation, which is then used to obtain improved segmentation results. The progress in applying AMs in this field suggests promising directions for future research, with potential for developing even more sophisticated models that can handle increasingly complex datasets and specific land cover challenges.

4) Transformers: Transformers are a type of neural network architecture initially developed for NLP tasks. They were conceived to handle sequential data, enable parallel processing, and ultimately address the “memory” limitations of RNNs. Transformers operate based on AMs, being the inaugural neural network model to be entirely founded on them without the need for recurrence. As shown in Fig. 26, the Transformer follows an encoder–decoder structure, where both components primarily consist of multihead attention modules, which in turn are built from multiple layers of self-attention mechanisms. In addition, transformers employ positional encoding, which provides information about the order of elements in a sequence, crucial for understanding context. Given the high performance demonstrated by Transformers, they have become a fundamental model in AI.

Transformers were soon adapted for tasks in CV, leading to the development of the Vision Transformer (ViT) [201]. The ViT was introduced to leverage the strengths of transformers for image analysis. Unlike traditional CNNs that rely on convolutions to capture local patterns, the ViT models images as sequences of patches. Each image is divided into fixed-size patches, which are then flattened and linearly embedded [202]. These embeddings are combined with positional encodings to retain spatial information, and the resulting sequence is fed to a standard Transformer encoder. The output of the Transformer encoder consists of a series of vectors, each corresponding to an input patch. In the original ViT (see Fig. 27), a special “classification token” is appended to the sequence of patch embeddings, and the corresponding output vector from the encoder is used as the image representation. This output vector is then passed through

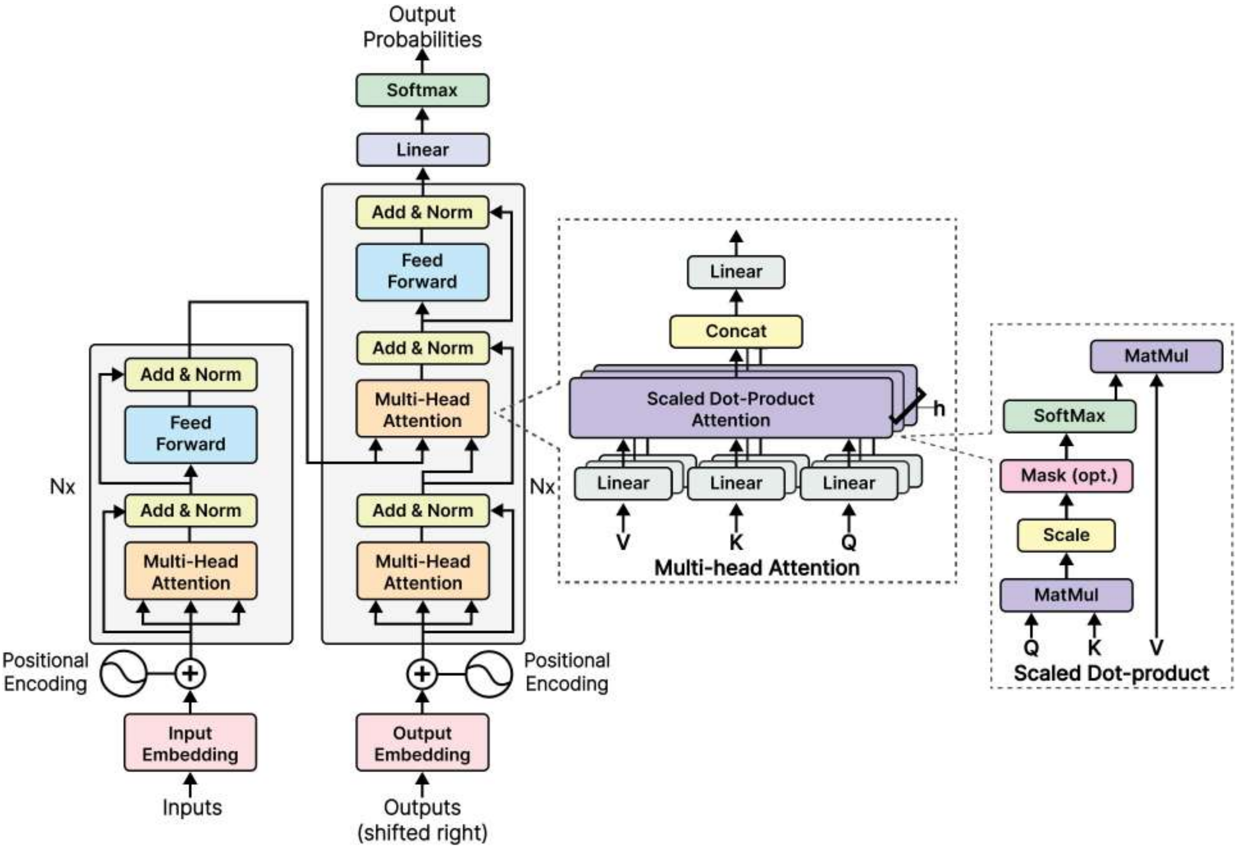


Fig. 26. Transformer architecture and its main components.

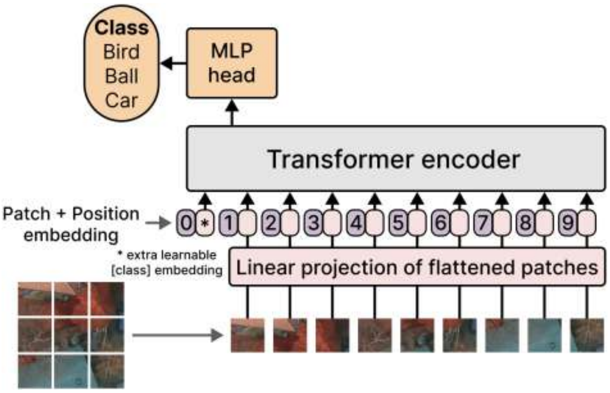


Fig. 27. ViT architecture.

a multilayer perceptron head for image classification. However, the flexible nature of the ViT architecture allows it to be adapted for other vision tasks such as object detection, image captioning, segmentation, among others.

In the field of LCC using MSI, the use of Transformers has found its natural application, given their effectiveness and ability to analyze images with high precision. Furthermore, Transformers for CV have continued to improve, with new variants being developed to extend their range of applications. Importantly, as is evident, Transformers inherently possess the capabilities of AMs, a topic previously discussed. Therefore, this group of works examines contributions that specifically use Transformers as an architecture, rather than just isolated AMs.

In this review, it is observed that Transformers are primarily used as feature extractors/encoders. This may be because transformers are highly effective at capturing complex relationships and long-range dependencies in the data, making them ideal for extracting detailed and contextual features of different types of land cover, which can then be classified. For example, the authors of [203], [204], and [205] use a Swin Transformer [206] as the sole and primary encoder within their LCC architectures; the first to process RG-NIR-DSM, and the last two for RG-NIR images. The Swin Transformer is a variant of the ViT that employs a shifted-window structure for reduced computational complexity and hierarchical feature extraction capabilities.

Other variants of ViT are also used, such as in [207], where the encoder is composed of four Mix Transformer [208] blocks that implement efficient self-attention instead of the standard self-attention used in ViT, reducing computational complexity when processing RG-NIR images. For the same type of image, in [209], the Pyramid ViT [210] is used as the encoder, which allows for multilevel feature extraction and uses a progressive shrinking pyramid to reduce the computations of large feature maps, along with spatial-reduction attention to further reduce resource consumption.

In other contributions, Transformers are used to complement other CNN-based encoders. This allows for the extraction of global semantic information with the help of the Transformer and the low-level contextual spatial information provided by CNNs, which can then be fused or processed to obtain rich features that guide better segmentation. In this group of works, Chen et al. [211] adopt a serial double coding structure using four ResNet50 blocks along with eight Transformer blocks to segment RG-NIR images. The work in [212] uses two parallel encoders, specifically a Swin Transformer and a ResNet34, to extract features from the same image (RG-NIR) simultaneously, with connections between the two branches that allow for feature fusion at different stages. Similarly, He et al. [213] use a Swin Transformer and link it with a UNet to form a parallel dual-encoder structure with a primary encoder, UNet, and an auxiliary encoder, Swin Transformer, for rich feature extraction from RG-NIR images, that will then be used by the original decoder of the U-shaped network for obtaining the segmentation result.

Some works use the Transformers not only for feature extraction but also experiment with their capabilities for the decoding phase. Liu et al. [214], for example, implement a fully Transformer-inspired approach, where a ViT functions as the encoder for feature extraction of RG-NIR images, and a Transformer decoder is designed to receive the features from the ViT, resize them to different levels, fuse them, and integrate both global and local semantic contextual information to make predictions. In [215], the focus is on using Transformers specifically for the decoding part using RG-NIR images. The work uses a ResNet50 as the encoder and develops a channel-spatial Transformer block that aggregates the local, channel, and global features extracted by the encoder, connected in a serial manner with a global cross-fusion module for the interactive fusion of the features, which are then used for prediction.

In summary, it is evident that the use of transformers has found significant applicability in this field. They are extensively used for encoding due to their ability to capture complex relationships within the data, often becoming the preferred choice over traditional CNNs. Moreover, there is still room to explore the capabilities of Transformers in decoding processes, as well as the potential development of more efficient variants. This is particularly important given that the inherent computational cost of Transformers is higher than that of CNNs. As research progresses, we can expect further innovations that leverage the strengths of Transformers while addressing their limitations, thereby enhancing their applicability in LCC.

5) Multiscale Techniques: Various approaches identified in the literature makes use of multiscale information to address MSSS for LCC. Multiscale techniques involve analyzing data at various spatial resolutions to capture details at different levels of granularity, as shown in Fig. 28. This allows for a more nuanced understanding of the land cover features, as both fine and coarse details can be considered, potentially enhancing the accuracy and robustness of segmentation models.

To begin, it has been identified that several works make use of atrous spatial pyramid pooling (ASPP) [216] within their architectures to leverage multiscale information. ASPP is a module that enables the resampling of a given feature layer at multiple rates, as shown in Fig. 29. In other words, it allows an image to be analyzed at different scales using different dilation rates, capturing more contextual information from it [217]. This is achieved by applying Atrous convolutions, also known as dilated convolutions [218], in parallel layers, each with its own unique sampling rate.

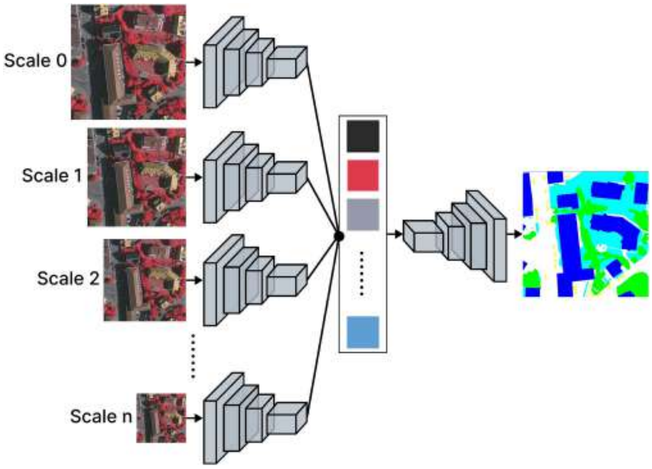


Fig. 28. Illustrative example of the use of multiscale information.

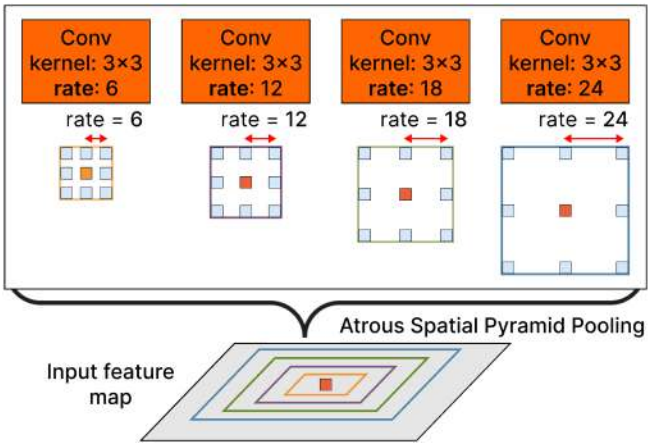

Conv kernel: 3×3 rate: 6
Conv kernel: 3×3 rate: 12
Conv kernel: 3×3 rate: 18
Conv kernel: 3×3 rate: 24
rate = 6
rate = 12
rate = 18
rate = 24
Atrous Spatial Pyramid Pooling
Input feature map

Fig. 29. Atrous spatial pyramid pooling.

The ASPP module is usually placed after the backbone or feature extractor to process the extracted features at multiple scales. Under the DeepLabV3 framework, Hou et al. [219] use a simple approach linking an ASPP module after a ResNet101 encoder and concatenating the features to obtain the segmentation map from RGB-NIR images. In [70], an ASPP module is also integrated after a ResNet101 backbone but opts to use a fully differentiable forest as the decoder for segmenting RG-NIR images. Similarly, Chen et al. [197] employ a ResNet101 as the backbone to process RGB-NIR and RG-NIR images and link it with an ASPP, but modify the latter by removing the global average pooling branch to reduce noise. Zhang et al. [220] use ResNet pretrained backbones (50 and 101) along with an ASPP module and additionally incorporate connections that link the encoder and the decoder through the ASPP to obtain additional features that enhance segmentation from RG-NIR-nDSM images.

In [17], an ASPP module is integrated in each skip connection of a UNet across four levels of encoding–decoding for segmenting RGB-NIR-SWIR images. Nuradili et al. [31] operate on the DeepLabV3 framework employing a ResNet50 backbone, which is connected to an ASPP to extract multiscale features from RGB-NIR-RE-TIR images, which are then decoded with the DeepLab decoder head. Similarly, Wang et al. [221] work within the DeepLabV3+ framework and RG-NIR images, integrating an ASPP module after a ResNet (50 and 101) encoder; however, they modify the decoder by using a tri-branch decoder. The work in [222] uses a multiscale context-feature-aggregation process based on ASPP. Two modules are added in the encoder stage to fuse features from multiscale receptive fields of RGB-NIR and DSM images, respectively, and obtain global multiscale context information to guide decoding. In [160], another variant of ASPP is used for handling multiscale information through the incorporation of depthwise atrous spatial pyramid pooling (DASPP) to process RGB-NIR images. DASPP implements atrous depthwise separable convolution to replace atrous convolution of ASPP, thereby lowering the computational cost while maintaining high performance.

Other approaches seek to extract multiscale information effectively for later use. For example, Liu et al. [156] incorporate a multiscale feature extraction module, utilizing differently scaled patches to gather both global and local image information from RG-NIR and RGB-NIR images. This division into large- and small-scale patches allows for an enhanced input that enriches the extraction process for multiscale features. For example, Li et al. [223] adopt an architecture with high-resolution resource extracting network [224] as the backbone for RGB-NIR and RG-NIR images, which integrates high- and low-resolution branches in parallel that decrease in resolution, progressing through four stages where each one produces an additional branch that interacts with the others. It then integrates a pyramid pooling module to extract contextual information at different scales, which are subsequently concatenated and used for prediction.

Moving forward, other works focus on multiscale information handling with approaches different from the use of ASPP. For instance, Zheng et al. [225] use four ResNet101 blocks as the backbone, each extracting information at different scales from RGB-NIR and RG-NIR images. These blocks are connected to four decoding phases through a module that performs weighting and summing of the features, thereby fusing the multiscale information. Similarly, Zhang et al. [85] use a UNet with a ResNet34 backbone as the baseline and add a multilevel aggregation module to encode contextual information across different scales and adaptively fuse them to integrate both global and local information from 13- and 18-band images. Cheng et al. [80], instead of simply concatenating features from different hierarchical layers of a CNN backbone, use a multiscale global context fusion module based on entropy to adaptively merge global features from RGB-NIR images. In addition, it employs a multilevel context integration module to capture local context information, such as dense areas of small objects. In a related development, Zhou et al. [226] incorporate a dense cross-decoder for multiscale dense fusion, exploiting rich semantic information in high-level features to guide and refine low-level features comprehensively from RG-NIR images.

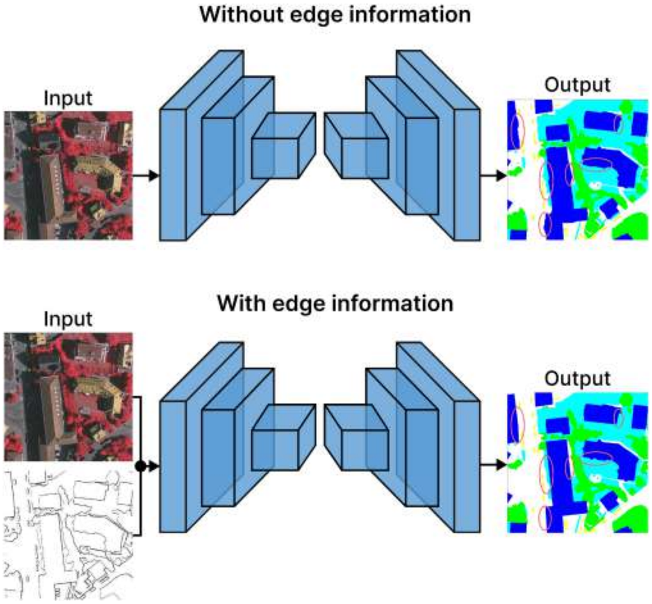


Fig. 30. Concept of the use of edges to improve LCC.

In summary, the use of multiscale information has emerged as a significant area of interest in the field of LCC using MSI. Given the inherent use of aerial imagery in this field, analyzing information at different levels can provide better context and detail, leading to more effective segmentation and classification of various land cover types. The reviewed works demonstrate diverse methods related to multiscale information, from extraction to fusion, showing that this is not only a recurrently explored field but also one with room for proposing new methods that enable a better and more efficient use of this type of information.

6) Use of Edge/Boundary Information: Another approach identified involves the utilization of edges, or boundaries, to enhance LCC. Edges can be defined as the points in a digital image where there is a significant change in intensity. This set of points forms boundaries that allow differentiating disjoint regions. Edges provide important information about the characteristics of an image, such as corners, lines, and curves; they also provide strong visual clues that can help in different pattern recognition processes.

In LCC, edges often correspond to transitions between different land cover types, providing information that could be useful for accurate classification. By leveraging edge data, these approaches aim to improve the delineation of distinct land cover classes, leading to more precise and reliable segmentation results and consequently providing better insights into the spatial distribution of the land surface. An illustration of the use of edges for LCC can be seen in Fig. 30.

To start with, in [192], a boundary-aware dual-stream network is introduced, featuring two components: a segmentation subnetwork and an edge detection branch. Both subnetworks are based on UNet, but the edge network incorporates fewer pooling layers. The decoders of the subnetworks are connected through a module that introduces boundary information from the edge subnetwork to the segmentation subnetwork, using a design similar to an AM but without the self-attention features. This approach enables more precise and well-defined segmentation from RGB-NIR and RG-NIR images. Similarly, Wang et al. [221] propose a tri-branch network that uses edge information to improve segmentation and estimate height from MSI RG-NIR imagery. The inputs are first processed by a ResNet (50 or 101) backbone, and then, those features are fed to each branch. Like the previous work, boundary information is shared at the decoder level through a boundary masking module, AM-like, that incorporates spatial boundary information into both the segmentation and height task branches to enhance their accuracy.

Other contributions utilize edge information not connected to the decoders but instead employ it in earlier stages of the segmentation process. For example, in [85], an encoder–decoder approach is used with three main components: a 3-D spectral information extraction module, a spatial information extraction module, and an edge information extraction module. The latter follows a siamese encoder structure with multiple layers that extract edge features and fuse them with the previous layer's features. The edge module shares weights with the spatial module, and the features obtained from this sharing are used alongside the spectral module's features for decoding and generating segmentation from MSI 13- and 18-band images.

In other instances, edge information is used as a guide to direct the segmentation process, aiming for better performance. In [227], an edge detection guidance module is used to improve segmentation from MSI images and their DSM information. Specifically, edge detection guidance is utilized to extract boundary features and use them as additional weights to assist the adaptive multimodal (RG-NIR+nDSM) fusion between high-level and low-level features of the inputs processed by ResNet50 blocks, ultimately decoding and obtaining segmentation maps. Similarly, in [228], an edge-aware module is incorporated within the spatial branch of a dual-branch encoder architecture, which upsamples, concatenates, and convolves spatial features to obtain edge information. This edge information, used as weights, guides the aggregation of semantic and spatial features, which are then decoded to generate segmentation maps from RG-NIR images. In similar fashion, in [229], a boundary guidance module is proposed, consisting of a simple structure formed by convolutions. This module combines low-level and high-level features from a Res2Net-50 backbone and assigns weights to these features based on the relevance of their semantic boundary information. Consequently, this generates features with rich semantic information linked to object boundaries from RGB-NIR and RG-NIR images, which are used along with spatial information to achieve segmentation with less ambiguity at the edges.

This group of works indicates that there is a significant current research avenue focused on leveraging edge information for MSSS. Specifically, in LCC, one of the challenges is precisely achieving an adequate differentiation of the different cover types and obtaining uniform and consistent edges between each of them, especially considering the type of imagery used and the perspective. Therefore, these approaches promise to enhance the ability of models to precisely delineate the boundaries between different land cover types, allowing for a more detailed and nuanced classification.

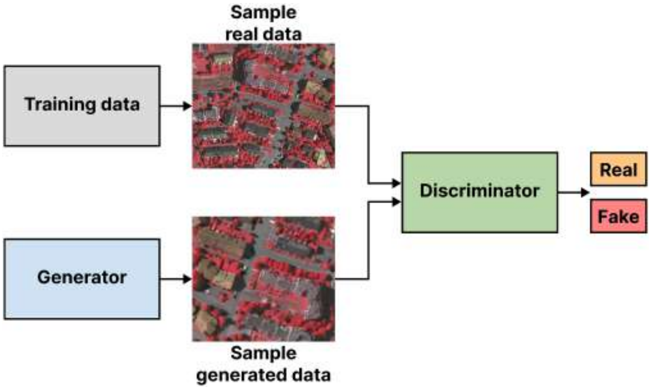


Fig. 31. Basic structure of a GAN.

7) Generative Approaches: In the field of LCC using MSI, several challenges can arise that complicate the segmentation process. These challenges include problems such as poor image resolution, limited data availability, and complexity in acquiring new images. Generative AI offers promising solutions to these challenges. Generative AI refers to models that can create new data samples similar to the original data. By generating synthetic data that closely resemble real data, these models can augment and diversify training datasets, making them more robust and comprehensive. Generative approaches can also enhance image resolution, extract meaningful features, and provide supplemental inputs that improve the accuracy of segmentation tasks.

Generative approaches have their foundations in methods such as autoencoders, which learn efficient codings of input data to reconstruct it accurately, and variational autoencoders (VAEs), which introduce a probabilistic approach to the encoding process to generate new data samples. However, the trend has shifted toward more advanced methods such as GANs. A GAN is a DL architecture that consists of at least two modules: a generator and a discriminator, as shown in Fig. 31. These modules are trained simultaneously through adversarial training, in which the generator creates synthetic samples following the distribution of the real data, and the discriminator evaluates the generated data against real. The goal is to adjust the architecture so that the generator is capable of creating high-quality synthetic data that are indistinguishable from real data by the discriminator.

Given the ability of generative techniques to create high-quality data, the literature reviewed shows that they are being used in the field of LCC using MSI to address certain limitations with the available data. For example, generative methods can be used to handle, improve, or correct the resolution of images. In [230], a GAN-based approach is employed for LCC using high-resolution RGB-NIR images reconstructed from low-resolution images. A progressive growing GAN [231] is used, trimming the generator and adding skip connections to improve network stability. The generator aims to produce a high-resolution remote sensing images that closely resemble the reals. These reconstructed images are then processed by

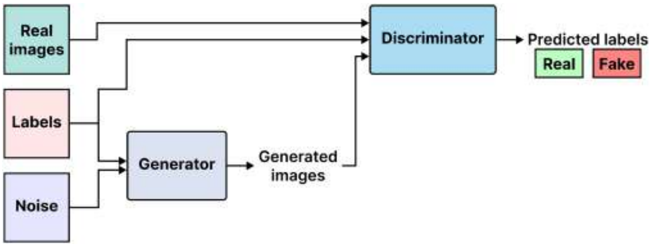


Fig. 32. CGAN structure.

classifiers to obtain segmentation maps. In [232], GANs are also used to handle image resolution, employing a multistage framework where each level consists of a UNet-based generator and a ResNet18-based discriminator. The network structure works by starting with a low-resolution MSI (combinations of RGB, NIR, NDVI, and nDSM) image and incrementally generating a high-resolution segmentation mask through each level. The initial generator takes the original image as input, while subsequent generators use the upsampled output from the preceding layer, progressively improving the precision and detail of segmentation.

Other approaches leverage generative techniques to augment the available data, allowing for the training of more robust segmentation models. In [196], an architecture with a generator–discriminator design is proposed. The generator, based on deep convolutional GAN, creates synthetic images from random noise through deconvolution layers to improve model robustness. The discriminator processes these images alongside real data, splitting them into different spectral bands and outputting class predictions and authenticity verification. In [233], a GAN is also used for data augmentation and consequently to improve model performance in LCC. This contribution introduces a conductor, which is an encoder–decoder network that learns different semantic embeddings and provides feedback to the generator for the synthesis of better MSI-like (RGB-NIR) images. Shi et al. [234] also utilize a GAN to augment the available data by using two very-high-resolution samples from different classes combined as input. This input is fed into the architecture that generates synthetic samples containing features from both original samples, thus creating images, RG-NIR-like, close to the classification boundary to form more robust datasets.

Furthermore, other GAN variants are also employed for MSI data generation. For example, in [235], a conditional GAN (CGAN) [236] is used to generate fake remote sensing MSI-like images from predefined ground truths used as the “condition.” A CGAN, shown in Fig. 32, is a type of GAN where the generation process is conditioned on additional information, such as class labels, allowing for more controlled data generation. The used CGAN is trained with pairs of real MSI (RG-NIR) images and their corresponding ground truths, and is used to increase the diversity of the training dataset that is subsequently utilized by CNN models for segmentation. Similarly, Sui et al. [237] apply a CGAN to augment training images, incorporating edge and boundary features. The generator used is UNet-based, while the discriminator is based on PatchGAN [238]. In this case, for training, the “condition” used includes the ground truth, and edge features, along with noise, to generate MSI (RG-NIR)-like images from this input.

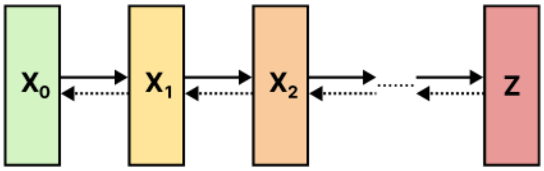


Fig. 33. Operation of a diffusion model by gradually adding noise and then reversing.

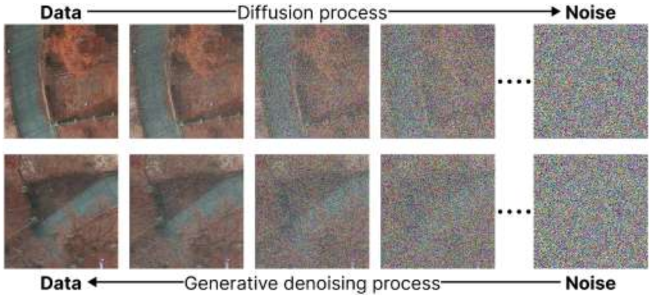


Fig. 34. Example of image generation through a diffusion process.

Generative techniques can be used to generate spectral information to complement existing datasets. For example, Paoletti et al. [239] generate DSMs using an architecture composed of two encoder–decoder VAE generators and two discriminators with a shared latent space. This architecture allows the generation of DSM from very high resolution optical images, through an image-to-image translation process. These DSMs can later be used as additional input to improve the performance of MSI segmentation, such as for RG-NIR images for LCC with which this approach was tested. Similarly, Costa et al. [240] use a CGAN to generate DSMs from very-high-resolution images. The framework begins by segmenting an input image with a SegNet, and the output is fed into a UNet-based generator to produce a DSM. The generated DSM is evaluated by the discriminator, which provides corrections based on the original DSM, training the generator to create accurate DSMs.

More recently, sophisticated generative approaches have been proposed, positioning themselves as potential tools for inclusion in LCC systems, such as diffusion models [241]. A diffusion model is a type of generative model that is trained to learn how to reverse a noise-adding process, as shown in Fig. 33. A diffusion process iteratively introduces noise, usually Gaussian noise, to the original data, causing it to gradually lose its initial characteristics. A diffusion model is trained to reconstruct the original data through an inverse diffusion process, allowing it to learn the data distribution and consequently gain the ability to generate new samples, as shown in the example of Fig. 34. These models have shown great promise in generating realistic and high-quality images and could be adapted for use in LCC using MSI.

In [242], for example, the use of diffusion models for the segmentation of aerial images is proposed. This approach begins with an initial segmentation generated as pure noise. The noise function is applied iteratively to diffuse the previously predicted segmentation. Then, the diffusion model uses a real image as a condition to remove the noise from the diffused segmentation. The resulting segmentation is compared with the ground truth to adjust the noise function recursively. This type of diffusion techniques promises to improve the robustness and accuracy of segmentation models using MSI, as demonstrated by the mentioned study, which showed successful aerial building segmentation from RG-NIR images.

As seen, generative techniques are highly useful for LCC tasks. In a field where the quantity and quality of images can be a limiting factor for developing robust models, generative techniques present themselves as the best option. They have been shown to address various tasks, from improving resolution to generating spectral information, demonstrating their broad applicability. In addition, there is still room for experimentation to generate other types of information or different spectral bands, further enhancing their potential in this domain. Moreover, novel generative techniques, such as diffusion modeling, offer new avenues for research and application in this field.

8) Independent Band/Spectra Processing: Typically, CV models are fed with input images without extensive preprocessing at the channel level. This is because most tasks and applications work with standard RGB images, and all tools are default-prepared to handle this configuration. When using MSI, the most straightforward approach is to process the images as a single entity by either expanding the number of channels in the input, stacking the spectral bands, or preliminarily fusing all the spectral bands into a single tensor. However, these approaches may not fully leverage the information provided by each band or spectral information. To address this, some studies have explored the idea of processing spectral bands separately, with the aim of improving the utilization of the information they provide. By separating the spectral bands or grouping them into specific subsets, as shown in the example in Fig. 35, it is possible to conduct a more targeted analysis of the unique information contained in each band, potentially leading to better classification and segmentation results.

To begin with, some approaches have opted to split different spectral bands. Xu et al. [243], for example, divide the input image into two modalities: RGB and NIR. To process this information, it adopts a dual-encoder and mono-decoder structure, where each spectral group is fed into its respective encoder for simultaneous feature extraction. The encoders have the same structure based on convolutional layers and pooling, sharing skip connections with the decoder, which receives the feature maps and processes them to obtain the prediction. Following a similar design, Wang et al. [244] utilize two CNN encoder networks to process an input divided into RGB and NIR, respectively. Each branch simultaneously extracts the features, which are then fused with shared weights and decoded. Similarly, Jiang et al. [245] also divide the input MSI into RGB and NIR. For processing, they adapt a UNet, duplicating its backbone to simultaneously extract features from each spectral group, which are then concatenated and upsampled to obtain the prediction. The same approach is followed in [246], where a UNet is modified to incorporate two separate input nodes to encode RGB and NIR, respectively. These inputs are then concatenated and decoded using the UNet decoder.

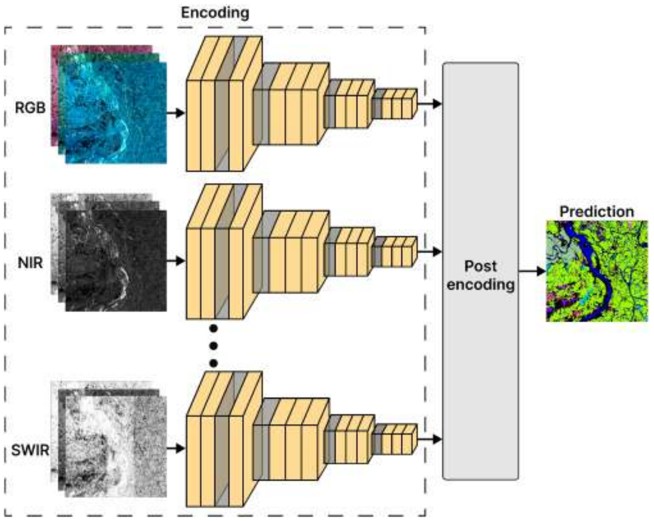


Fig. 35. Illustrative example of a encoding process for an MSI image following the approach of band grouping into different spectra.

Other approaches handle more than four channels in their architectures. To begin with, Tao et al. [23] divide the bands of an MSI into two groups, visible and invisible light, for simultaneous feature extraction. Specifically, they propose two configurations: one that processes RGB and NIR bands (four channels in total) and another that processes RGB and NIR+spectral indices (six channels in total). To process this information, they use an approach with two twin ResNet50-based encoders, each receiving a spectral group (visible and invisible). The first backbone extracts color features, texture features, and spatial relationships, while the second extracts spectral features, which are then decoded to obtain the segmentation. In [247], an RGB-NIR image is divided into two combinations, i.e., RGB and GB-NIR, to obtain two inputs, totaling six channels. Each of these inputs is fed into its respective ResNet-based backbone for feature extraction. Each feature map is then upsampled and concatenated using two horizontal connections to obtain the prediction. Zhang and Yang [196] expand the range of bands used, handling ten-channel images. These images are first split into four different spectral groups based on their optical characteristics: RGB, NIR, RE, and SWIR. Each group is then encoded by a respective CNN-based backbone, each with the same structure, and the features are concatenated and decoded to obtain the classification.

Certain methods opt to process MSI data in groups, also including other types of spectral information in addition to channels. For example, Li et al. [248] implement a dual-branch approach to process an RG-NIR image on the one hand and its

DSM on the other hand. To do this, two feature extractors are used: one based on Xception for the RG-NIR part, and another branch with convolutional layers, separable convolutions, and inner residual blocks to process the DSM. These features are then concatenated and decoded to obtain the segmentation. A similar approach is used in [222], which utilizes a configuration of two twin ResNet50 backbones to process an RG-NIR image and its DSM, extract their features, subsequently fuse them, and finally decode them. The same applies in [249], with the difference of using a ResNet101 to form two extractors responsible for processing MSI images (RGB-NIR or RG-NIR) and their DSMs, respectively. In a slightly different approach, Zhou et al. [250] varies in the use of spectral information by addressing the processing of RG-NIR images and their nDSM. To achieve this, two ResNet50 backbones are applied to extract features, which are then fused, refined, and used to generate predictions. In a similar approach, Zhou et al. [251] also processes RG-NIR and nDSM information using two encoders; in this case, a ResNet50 is used for RG-NIR and a MobileNetV2 for the nDSM to generate features, which are then subsequently decoded.

In summary, grouping or splitting the spectral information of MSI images for separate processing is a promising approach that can improve data utilization and achieve better segmentation results. The reviewed works in this group suggest that this approach is primarily used in the encoding or feature extraction stage to capture detailed and specific information from each spectral group. This can potentially contribute to more effective combination of features, more precise decoding, and, consequently, better classification of different land cover types.

IX. Discussion and Outcomes

In this section, we present the key research outcomes derived from our review. We aim to highlight prevalent trends, approaches, and emergent themes that characterize the current state of research. These outcomes not only reflect the advancements within the domain but also offer insights into potential future directions, enabling us to infer underlying patterns and priorities that are shaping the evolution of LCC. In addition, Table XIV presents a summary of all the contributions considered in this review that provides additional insights.

A. *Comments on MSSS Methods for LCC*

The landscape of MSSS methods for LCC exhibits a clear trend toward the utilization of AI techniques. Interestingly, this review did not uncover any studies employing traditional image processing methods for this task, underscoring a dominant preference for AI-driven approaches in recent years. A deeper analysis reveals a particular focus on DL rather than ML. While our review did identify a number of studies employing ML techniques, their application was predominantly oriented toward specific case studies rather than the development of new methods or approaches based on these techniques. This trend suggests that ML plays a complementary role, rather than a leading one, in advancing MSSS for LCC.

It is noteworthy that ML methods are applied in both PBIA and OBIA approaches. Moreover, the MRS algorithm, primarily utilized through the eCognition software,[21] was the most common method for object delineation prior to classification. Furthermore, for the semantic classification of pixels, SVM and RF emerged as the most frequently applied techniques in the corpus of reviewed studies. They are also the methods that report the best numbers, with figures above 90% in several cases [165], [167], [169], [170]. Although less prevalent, the continued use of these ML methods reveals their robustness and capability to meet the demands of specific application fields, consistently delivering acceptable performance.

Regarding the DL methods in our review, we have identified eight distinct groups of contributions within this category. The first two groups comprise studies that use preexisting methods and architectures directly and with slight modifications, respectively. The works in these two groups are notable for being mostly focused on case studies or analysis of specific areas rather than the development of new robust methodologies. However, the performance these methods deliver is commendable, with accuracy rates frequently around 80%. Moreover, among the architectures used in these groups, while SegNet and ResNet are noteworthy for their contributions, a clear preference emerges for UNet, in most of the works of these groups [5], [119], [172], [173], [174], [183], [184], [185], as a particularly popular choice both for direct application and for the incorporation of small modifications. This trend toward UNet, in comparison to others, likely stems from its proven efficiency and effectiveness in complex segmentation tasks, suggesting a strategic move within the research community to leverage reliable and established models to navigate the intricacies of LCC using MSI. The remaining groups, following these two, transition from contributions primarily focused on case studies to works that propose new methodologies.

The next two groups within DL-based methods share similarities as they involve AMs and Transformers, respectively. Regarding AMs, the studies employing these techniques highlight that AMs can be applied with different approaches, such as channel level or spatial level, which allows for great versatility and even the combination or stacking of several mechanisms within the same framework. Notably, many works [190], [193], [194], [195] base their proposed mechanisms on the SE network. In addition, the performance reported by these studies consistently remains in the range of 90% accuracy. The works using Transformers inherently leverage the capabilities of visual attention but at a more complex level. This group notably uses Transformers primarily as feature extractors. Despite employing various types of Transformers, the Swin Transformer stands out slightly above the others due to its hierarchical extraction capabilities [203], [204], [205], [212], [213]; furthermore, it is also combined with other encoders for effective feature extraction. In terms of performance, the contributions that use Transformers demonstrate notable results, usually in the range of 90% or higher.

[21] [Online]. Available: https://geospatial.trimble.com/en/products/software/trimble-ecognition

TABLE XIV
SUMMARY OF REVIEWED WORKS DETAILING SPECTRUM, DATA USED, AND KEY PERFORMANCE METRICS

Method	Image modality	Dataset/Data source	OA (%)	mPA (%)	Kappa	mIoU (%)	mF1 (%)	Ref.
RF	Eight bands	WorldView-2	91.9		0.91			[165]
SVM			87.3		0.85			
SVM	10 (MS) + SAR	Sentinel-1 + Sentinel-2	92.87		0.892			[167]
RF			91.92		0.879			
KNN			95.12		0.926			
SVM	10 (MS) + SAR + NDVI	Sentinel-1 + Sentinel-2	91.29		0.869			
RF			91.92		0.879			
KNN			95.12		0.926			
PCA + RF		WorldView-2	92.53		0.890			[168]
PCA + KNN			80.67		0.700			
Maximum likelihood	RGB-NIR	UAV-captured	73.68		0.663			[45]
KNN	RGB-NIR-NDVI-NDWI-MSAVI2	UAV-captured	94.90		0.934			
SVM	R-G-NIR	Sentinel-2	95.00		0.922			[169]
RF			95.00		0.920			
CART	RGB-NIR-SWIR-NDVI-MNDWI-NDBI	Sentinel-2	96.25		0.940			[170]
SVM			97.00		0.950			
RF			98.68		0.970			
UNet	RGB-NIR	GID	82.27		0.7721	64.05		[172]
UNet	RGB-NIR	GID	82.27		0.772	64.05		[173]
UNet	RGB-NIR	Gaofen-1	91.72					[5]
	RGB		91.21					
UNet	RGB-NIR	GID	84.8		0.83			[119]
UNet	Eight bands	Worldview-3				51.00		[174]
PSPNet						42.00		
SegNet	RGB-NIR-SWIR	Sentinel-2	85.00				84.00	[177]
SegNet	RG-NIR	Vaihingen	86.09		0.804			[179]
ResNet152			87.75		0.804			
Res-Seg-net	RGB-NIR	RIT-18	70.00					[182]
Res-UNet + inception	RGB-NIR	GaoFen-2	83.1					[183]
	GB, NIR		81.4					
	RB, NIR		81.9					
	RG, NIR		82.3					
	RGB		78.8					
Deep UNet	RGB-NIR	RIT-18	90.6			89.51		[184]
UNet 3D	RGB-NIR	RIT-18	90.698					[185]
NDVI-RSU-Net	RGB-NIR-NDVI	GID		88.36		71.18		[190]
	RGB-NIR			87.24		67.69		
	RGB			87.95		66.35		
MFAGNet	RG-NIR	Potsdam	9.71	87.24	0.8787	79.42		[156]
	RG-NIR	Vaihingen	86.65	83.35	0.8237	73.40		
	RGB-NIR	GID	84.10	79.79	0.7980	69.05		
ARLNet	RG-NIR	Potsdam	90.9					[193]
	RG-NIR	Vaihingen	90.5					
BLASeNet	RGB-NIR	Potsdam	79.64			62.90	75.47	[194]
	Seven bands	Qinghai	88.39			77.71	86.98	
LANet	RG-NIR-DSM	Potsdam	91.95				90.84	[195]
	RG-NIR-DSM	Vaihingen	89.83				88.09	

TABLE XIV
(CONTINUED)

Method	**Image modality**	**Dataset/Data source**	**OA** (%)	**mPA** (%)	**Kappa**	**mIoU** (%)	**mF1** (%)	**Ref.**
Ad-hoc	RGB-NIR	Potsdam	88.20			70.32		[197]
	RG-NIR	Vaihingen	87.12			73.46		
MPFFNet	RGB-NIR	GID	87.33			77.83		[160]
HMANet	RG-NIR	Vaihingen	91.44			83.49		[198]
SegVAN	RGB-NIR	Potsdam	91.68			87.29	93.10	[199]
	RG-NIR	Vaihingen	91.49			84.03	91.20	
MANet	RG-NIR	Vaihingen	90.96			82.71	90.41	[200]
BAM-UNet	RGB-NIR-DSM	Potsdam	89.13				88.59	[192]
	RG-NIR-DSM	Vaihingen	89.75				89.10	
BAMTL	RG-NIR	Vaihingen	88.40				86.90	[221]
ESSANet	13 bands	SEN12MS	89.07	83.26	0.8637	73.31		[85]
	18 bands	Urban Semantic 3D	92.28	91.46	0.8529	85.69		
EDGNet	RG-NIR-nDSM	Postdam		84.60		75.34		[227]
	RG-NIR-nDSM	Vaihingen		89.48		81.15		
EIGNet	RG-NIR	Vaihingen	91.21			83.05		[228]
BGFNet	RGB-NIR	Postdam	92.18			86.94	92.90	[229]
	RG-NIR	Vaihingen	90.57			81.65	89.85	
PGGAN-MS	RG-NIR	Gaofen-1 + Sentinel-2	98.93					[230]
Multi-level GAN	RGB-NIR-NDVI-nDSM	Potsdam	91.8				88.8	[232]
	RG-NIR-NIR-nDSM	Vaihingen	89.9				89.0	
Ad-hoc	RGB-NIR-RE-SWIR	Sentinel-2	98.77					[196]
CSEBGAN	RGB-NIR	GID	94.10			67.50		[233]
IGAN	RG-NIR	Vaihingen	81.23		75.06			[234]
CGAN-TSIM	RG-NIR	Vaihingen	75.85					[235]
ECGAN	RG-NIR	Vaihingen	75.00		67.20	53.7		[237]
U-IMG2DSM	RG-NIR-DSM	Potsdam	70.95		60.66			[239]
DeepLabV3 modified	RGB-NIR	GID				61.51		[219]
SDNF	RG-NIR	Vaihingen	92.6			83.9	91.1	[70]
JSH-Net (ResNet50)	RG-NIR-nDSM	Vaihingen	90.00			82.30	90.05	[220]
JSH-Net (ResNet50)	RG-NIR-nDSM	Potsdam	90.00			84.54	91.41	
JSH-Net (ResNet101)	RG-NIR-nDSM	Vaihingen	90.17			82.63	90.26	
JSH-Net (ResNet101)	RG-NIR-nDSM	Potsdam	90.05			84.60	91.48	
ASPP-UNet	RGB-NIR-SWIR	Landsat-5	78.82		0.7551			[17]
PCA + DeepLabV3	RGB-NIR-RE-TIR	UAV-captured		92.8		73.5		[31]
DASFNet	RG-NIR-nDSM	Potsdam		85.17		76.13		[222]
		Vaihingen		89.37		82.08		
HrreNet	RGB-NIR	QERED				73.05		[223]
	RG-NIR	Vaihingen				80.87		
GAMNet	RG-NIR	Vaihingen	91.7				93.1	[225]
	RG-NIR	Potsdam	91.3				90.6	

TABLE XIV
(CONTINUED)

Method	Image modality	Dataset/Data source	OA (%)	mPA (%)	Kappa	mIoU (%)	mF1 (%)	Ref.
MLCC	RGB-NIR	GID	86.44		0.861		93.72	[80]
GAGNet-S*	RG-NIR	Vaihingen		88.85		80.84		[226]
SwinB-CNN	RG-NIR-DSM	Vaihingen	91.4					[203]
Swin-S+DCFAM	RG-NIR	Vaihingen	91.63			83.22	90.71	[204]
STDSNet	RG-NIR	Vaihingen				81.77	89.81	[205]
2DSegFormer	RG-NIR	Vaihingen	88.85			77.49	87.08	[207]
PVT	RG-NIR	Vaihingen	88.87			71.79	82.05	[209]
SRCBTFusion-Net	RG-NIR	Vaihingen				76.27	86.26	[211]
STransFuse	RG-NIR	Vaihingen	86.07			66.66	78.67	[212]
ST-UNet	RG-NIR	Vaihingen				70.23	82.15	[213]
GLOTS	RG-NIR	Vaihingen		77.81		70.13		[214]
HAFNet	RG-NIR	Vaihingen	90.29			76.37	85.93	[215]
MUFNet	RGB-NIR	GID	78.10			44.60		[243]
DBED	RGB-NIR	GID	81.82	72.07		59.21		[244]
UNet+DCAM	RGB-NIR	GID	87.80		0.671	50.30		[245]
SiU-Net	RGB-NIR	Sentinel-2	93.00				79.70	[246]
MSNet (6-c)	RGB-NIR-NDVI-DSM	GID	93.85			84.23		[23]
MSNet (4-c)	RGB-NIR	GID	93.49			84.34		
MSNet (6-c)	RGB-NIR, NDVI, DSM	Potsdam	95.64			89.24		
MSNet (4-c)	RGB-NIR	Potsdam	95.02			88.02		
DS-FusionNet	RGB-NIR	Gaofen-2	70.3			43.6		[247]
DSPCANet	RG-NIR-DSM	Potsdam Vaihingen	90.13 87.32			77.66 72.56	87.19 84.46	[248]
MSDFM	RG-NIR-DSM	Potsdam Vaihingen	92.86 90.04				92.72 90.11	[249]
CIMFNet	RG-NIR-nDSM	Potsdam Vaihingen		84.80 89.31		75.67 81.44		[250]
MSTNet-KD	RG-NIR-nDSM	Potsdam Vaihingen		85.57 89.75		75.71 81.56		[251]

Results obtained using RGB or other combination of bands not within the scope of this review were excluded.

This review also notes a consistent trend in leveraging multiscale information for LCC. Specifically, the use of ASPP modules has become prominent in many current contributions [17], [31], [70], [160], [197], [219], [220], [221], [222], either through their direct integration or by proposing modified versions. Other works propose their own modules for the extraction or fusion of multiscale information. In any case, the performance of these approaches, like the previous groups, remains around 90%. The use of edge/boundary information has also stood out among the reviewed works. These studies primarily focus on leveraging edge extraction information to improve segmentation and achieve better differentiation between various land cover types. Some approaches work at the encoder level [85], others at the decoder level [192], [221], and some use edges as a guide [227], [228], [229]. In any case, all of them utilize edge information along with spatial and spectral data for their methodologies. The performance of this group again positions itself around the 90% threshold.

Another notable aspect is the inclusion of generative techniques. These techniques address various problems with the available data for training models and are even used as integrated components in segmentation frameworks. The contributions in this group use generative techniques to tackle issues such as poor image resolution [230], [232] and the generation of synthetic data to compensate for the scarcity of training data. This generation ranges from MSI-like images [233], [234], [235], [237] to other types of spectral information such as DSMs [239], [240]. Regarding specific architectures, the use of conditional approaches like CGAN stands out, but other methods like diffusion models also appear, which have not yet been fully explored for LCC tasks, specifically using MSI, and represent an open area of research.

Finally, the independent utilization of different types of information from MSI images emerges as a coherent approach, aimed at leveraging the data in the best possible way. According to the reviewed literature, these contributions primarily focus on the feature extraction phase using one or more encoders for determined groupings, usually groups of visible and nonvisible bands [23], [196], [243], [244], [245], [246], [247]. In addition, this methodology is also observed to exploit channel information on the one hand and other types of spectral information such as DSMs and nDSMs on the other hand [222], [248], [249], [250], [251]. All these approaches are designed for more efficient and effective use of MS information, rather than simply stacking channels, which could lead to the loss of important information. In terms of performance, there is a slightly varied behavior with approaches closer to the 80% accuracy threshold, while others exceed 90%. However, these types of contributions present themselves as very promising approaches due to their design, and they open multiple research avenues for developing frameworks that fully leverage the different spectral information in the most effective way.

B. Remarks on the Use of Nonvisible Bands for LCC

In this review, certain trends have been identified regarding the use of nonvisible bands for the task of LCC. As previously mentioned, our focus was solely on MSI of up to 36 bands. Within this domain, the NIR band has been identified as the most frequently used to enhance classification performance. This band was used in all the studies reviewed, often accompanied by visible bands or others within the spectrum. This widespread adoption suggests a recognition of the NIR band's unique capabilities in providing critical information that is not captured by visible light. Enhancing features such as vegetation vigor, moisture content, and other environmental markers, the NIR band significantly contributes to the accuracy and depth of classification outcomes. Its consistent use underscores the essential role that NIR and other nonvisible bands play in enriching the data inputs for more effective and precise LCC. Furthermore, many of the methodologies within this group are centered around and evaluate their new methods using datasets that specifically include the NIR band. This is further evidenced by the fact that many studies, aiming to avoid increasing the number of channels introduced into the network, prefer to discard other bands from the visible spectrum and retain the NIR band. Specifically, it was found that the blue band was most frequently omitted, with the RG-NIR triplet emerging as one of the most commonly used combinations, as can be seen in Table XIV. This selective use of spectral bands underscores the strategic prioritization of NIR due to its valuable contributions to enhancing classification accuracy, while simultaneously streamlining the input data to make the computational process more efficient. This focus highlights the importance placed on the NIR band in current research, recognizing its vital role in providing enhanced analytical capabilities that significantly improve the accuracy of LCC models.

Besides the NIR band, other bands, such as the SWIR, RE, and Thermal, were also explored, albeit less frequently according to the findings of our review. These bands, while not as predominantly utilized as NIR, still offer unique advantages for specific applications within LCC. For instance, the SWIR band is particularly effective in penetrating haze, fog, and smoke and in differentiating between moisture levels of soil and vegetation. Similarly, the RE band is known for its sensitivity to changes in chlorophyll content, making it useful for monitoring vegetation health. The Thermal band, on the other hand, is invaluable for assessing temperature variations across landscapes, which can be indicative of water stress in plants or the presence of specific materials. Although these bands were not as commonly employed as NIR, their selective use in certain studies not only demonstrates their potential to provide supplementary information but also indicates a growing integration into the field.

Similarly, the incorporation of additional information, such as spectral indices and DSM, was also explored. Several studies in this review integrated this additional information within their contributions, even designating a specific branch for its processing, as analyzed in previous sections. The inclusion of this type of spectral information, as well as the methods for grouping, or combining them, constitutes room for future research parallel to the improvement of methods for processing them.

C. Insights Into Datasets and Image Acquisition for LCC

The trend and necessity of using aerial perspectives in LCC make image acquisition unique and costly. Unlike other fields where a common camera suffices, LCC requires specialized sensors, adding complexity and expense. This impacts the logistical and financial aspects of research and influences the accessibility of conducting advanced LCC studies, potentially limiting them to well-funded projects or institutions with the necessary technological infrastructure. Fortunately, technological advancements in recent years have somewhat simplified the complexity of acquiring the types of images used in LCC. As highlighted in this review, there are now accessible satellite services that often provide, sometimes free of charge, a wide range of tools for downloading and even preprocessing MSI for various tasks. This accessibility has significantly propelled this area of research, enabling the development of mapping studies across diverse regions and areas around the world that address a wide variety of goals, including the mapping of mining zones, urban population analysis, climate analysis, and more. This democratization of data access has opened new avenues for researchers and practitioners across the globe, enhancing the potential for innovative studies and comprehensive environmental monitoring.

Moreover, the availability of imaging services also fuel the creation of benchmark datasets within this domain. Such datasets are a crucial tool in the development of new methodologies and architectures, as they establish a benchmark against which new developments can be compared. In this review, we have identified several benchmark datasets that have become standards for the development and evaluation of new models, many of which are constructed from images provided by satellite services, such as the GID dataset. This availability not only streamlines research processes but also ensures a level of consistency and comparability across studies, enhancing the overall robustness and reliability of advancements in the field. This field not only benefits from well-established datasets, but also continues to see the development of newer, more comprehensive datasets that offer greater richness and complexity. For instance, the Catalonia Multiresolution Land Cover Dataset [252] is a recent addition that encompasses 41 distinct land cover classes, providing a detailed and diverse set of data for advanced analysis. Similarly, the Five-Billion-Pixels dataset [253] represents an expansion of the GID dataset, increasing from 15 to 24 classes, thus offering a more detailed and extensive resource for model development. These are examples of how technology provides solutions that enable access to increasingly better data, which in turn supports the development of more sophisticated and effective models. These developments not only enrich the resources available to researchers but also enhance the scope and depth of potential studies, driving forward the capabilities of LCC techniques.

A small subset of studies within this review has adopted the use of drones equipped with MS cameras to capture their own datasets. This adoption highlights a burgeoning integration of drone technology in the field of LCC, potentially simplifying data acquisition processes. However, the limited number of such studies could be attributed to the high costs associated with MSI equipment. Moreover, the operational complexities and the inherent limitations of drones, such as reduced coverage area compared to satellites, further explain their sparse utilization. Despite the advantages of flexibility and accessibility that drones offer, the substantial costs, operational challenges, and coverage limitations suggest that satellite imagery and the use of established benchmarks are likely to remain predominant methods of image capture in LCC for the foreseeable future.

Generative AI also emerges as a valuable tool for addressing data availability limitations in LCC. These methods enable the synthesis of high-fidelity, diverse, and extensive training samples that enhance dataset quality and breadth. For instance, GANs are used to generate synthetic MS images similar to those from Sentinel [254] or SAR imagery [255], effectively expanding the dataset while maintaining high relevance and applicability to real-world scenarios. In addition, diffusion models, like Stable Diffusion [256], also demonstrate potential here for dataset generation. Like GANs, diffusion models excel in generating high-quality synthetic imagery but with a focus on even more nuanced control over the generation process. Their ability to create detailed and varied synthetic imagery supports the expansion of data resources, facilitating more comprehensive and nuanced analyses. This not only enhances data availability, particularly in scenarios where traditional acquisition methods are limited, but also fosters the development of more effective and accurate classification techniques, paving the way for innovative solutions in the face of evolving challenges in LCC.

X. Review Limitations

One of the limitations of this review stems from the specificity of the search query used. By focusing on the term “land cover” for data extraction, there is a possibility that relevant studies that do not explicitly differentiate between “land cover” and “land use” may have been inadvertently excluded. This distinction is often nuanced in the literature, and the overlap between these two terms can vary significantly between studies. As a result, our search criteria might have limited the scope of included studies, potentially omitting valuable research that addresses similar themes under a slightly different terminology or focus.

In addition, while this review intentionally focuses on the most recent advancements, it inherently limits its scope to studies published within the last few years. This temporal focus was chosen under the assumption that earlier contributions have been well documented in previous reviews. Although this approach allows for a concentrated analysis of current trends and technologies, it may omit historical perspectives or foundational studies that are still relevant today. We believe that focusing on recent developments provides the most value given the rapid pace of technological progress in this field, yet we acknowledge that this choice narrows the breadth of historical context considered.

XI. Conclusion and Future Work

This work assesses the latest trends and methodologies in MSSS for LCC. The objective is to present an up-to-date overview that reflects the current state of research, spanning various methodologies, image acquisition strategies, preprocessing techniques, and evaluation criteria. In addition, it outlines potential future directions that could further enhance the accuracy and efficiency of land cover mapping, supporting the ongoing evolution of the field.

In the analyzed methodologies, there is a discernible trend toward the use of advanced DL technologies, which predominate over traditional ML approaches. ML is particularly recurrent in studies focused on analysis or case studies, while DL is geared toward the development of new methodologies. Within DL, approaches based on AMs, Transformers, edge-based, multiscale, generative, and independent spectral information processing are particularly notable, providing contributions capable of delivering remarkable performance.

Regarding the use of spectral bands, this review identifies the NIR band as the most frequently utilized beyond the classic RGB. In addition, there are indications of the adoption of other bands and spectra, such as RE, Thermal, and SWIR, suggesting an expansion of studies to include a greater number of bands. This trend highlights a growing interest in exploiting a broader spectrum of data to enhance the accuracy and comprehensiveness of LCC. Although research using RGB images is foundational, what is shown in this review highlights a growing interest in exploiting a broader spectrum of data to enhance the accuracy and comprehensiveness of LCC.

Regarding datasets and image acquisition, two dominant approaches are evident: the acquisition of satellite images, primarily for specific studies, and the use of benchmark datasets. Currently, there is a significant presence of satellite services that facilitate access to and use of imagery, along with several solid benchmark datasets that continue to support the development of new methodologies. The future looks promising as the ease of access to these services not only enables data acquisition for specific studies but also supports the creation of new benchmarks. In addition, there is a burgeoning incursion of drone systems equipped with cameras for data capture. While these systems are still limited and generally less refined compared to traditional methods, their use is beginning to gain traction in the field. Moreover, GANs and diffusion models are emerging as powerful methods for data generation, presenting strong prospects for both the present and future of LCC.

As a direction for future work, this review could be expanded to cover additional domains such as HSI and land use analysis. Exploring hyperspectral data could deepen the understanding of LCC by utilizing its finer spectral resolution to capture subtle differences in materials. Future research could focus on comparing methodologies across these domains, identifying the strengths and limitations of each approach in various contexts.

References

[1] A. H. Chughtai, H. Abbasi, and I. R. Karas, "A review on change detection method and accuracy assessment for land use land cover," *Remote Sens. Appl.: Soc. Environ.*, vol. 22, 2021, Art. no. 100482.

[2] J. Wang, M. Bretz, M. A. A. Dewan, and M. A. Delavar, "Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects," *Sci. Total Environ.*, vol. 822, 2022, Art. no. 153559.

[3] Q. Lei, H. Jin, J. Lee, and J. Zhong, "Land use and land cover change simulation enhanced by asynchronous communicating cellular automata," *Theor. Comput. Sci.*, vol. 985, 2024, Art. no. 114331.

[4] A. Ghorbanian, M. Kakooei, M. Amani, S. Mahdavi, A. Mohammadzadeh, and M. Hasanlou, "Improved land cover map of Iran using sentinel imagery within Google Earth Engine and a novel automatic workflow for land cover classification using migrated training samples," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 276–288, 2020.

[5] M.-J. Li et al., "Classification of surface natural resources based on U-NET and GF-1 satellite images," in *Proc. 17th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process.*, 2020, pp. 179–182.

[6] B. Chai and P. Li, "An ensemble method for monitoring land cover changes in urban areas using dense landsat time series data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 29–42, 2023.

[7] S. H. Molla and Rukhsana, "Mapping spatial dynamicity of cropping pattern and long-term surveillance of land-use/land-cover alterations in the Indian Sundarban region," *Arabian J. Geosci.*, vol. 16, no. 6, pp. 1–20, 2023.

[8] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[9] I. Papoutsis, N. I. Bountos, A. Zavras, D. Michail, and C. Tryfonopoulos, "Benchmarking and scaling of deep learning models for land cover image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 250–268, 2023.

[10] L. Ramos, E. Casas, C. Romero, F. Rivas-Echeverría, and M. E. Morocho-Cayamcela, "A study of ConvNeXt architectures for enhanced image captioning," *IEEE Access*, vol. 12, pp. 13711–13728, 2024.

[11] M. Peng et al., "Crop monitoring using remote sensing land use and land change data: Comparative analysis of deep learning methods using pre-trained CNN models," *Big Data Res.*, vol. 36, 2024, Art. no. 100448.

[12] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self- and unsupervised learning for image classification," *IEEE Access*, vol. 9, pp. 82146–82168, 2021.

[13] R. Jiao et al., "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," *Comput. Biol. Med.*, vol. 169, 2024, Art. no. 107840.

[14] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, Jun. 2020.

[15] M. K. Kar, M. K. Nath, and D. R. Neog, "A review on progress in semantic image segmentation and its application to medical images," *SN Comput. Sci.*, vol. 2, no. 5, Jul. 2021, Art. no. 397.

[16] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[17] W. Zhang, P. Tang, and L. Zhao, "Fast and accurate land-cover classification on medium-resolution remote-sensing images using segmentation models," *Int. J. Remote Sens.*, vol. 42, no. 9, pp. 3277–3301, Feb. 2021.

[18] T. T. Nguyen et al., "Monitoring agriculture areas with satellite images and deep learning," *Appl. Soft Comput.*, vol. 95, 2020, Art. no. 106565.

[19] Z. Kütük and G. Algan, "Semantic segmentation for thermal images: A comparative survey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 285–294.

[20] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5558–5565, Oct. 2020.

[21] W. Ji et al., "Multispectral video semantic segmentation: A benchmark dataset and baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1094–1104.

[22] C. Wang, C. Wang, W. Li, and H. Wang, "A brief survey on RGB-D semantic segmentation using deep learning," *Displays*, vol. 70, 2021, Art. no. 102080.

[23] C. Tao et al., "MSNet: Multispectral semantic segmentation network for remote sensing images," *GIScience Remote Sens.*, vol. 59, no. 1, pp. 1177–1198, Aug. 2022.

[24] N. E. Shaik et al., "Longwave infrared multispectral image sensor system using aluminum-germanium plasmonic filter arrays," *Nano Res.*, vol. 16, no. 7, pp. 10018–10025, May 2023.

[25] S. Kottner, M. M. Schulz, F. Berger, M. Thali, and D. Gascho, "Beyond the visible spectrum—Applying 3D multispectral full-body imaging to the VirtoScan system," *Forensic Sci., Med. Pathol.*, vol. 17, no. 4, pp. 565–576, Sep. 2021.

[26] P. Braga et al., "Vegetation indices and NIR-SWIR spectral bands as a phenotyping tool for water status determination in soybean," *Precis. Agriculture*, vol. 22, no. 1, pp. 249–266, Jul. 2020.

[27] D. Nath, R. Laik, V. S. Meena, B. Pramanick, and S. K. Singh, "Can mid-infrared (mid-IR) spectroscopy evaluate soil conditions by predicting soil biological properties?," *Soil Secur.*, vol. 4, 2021, Art. no. 100008.

[28] J. Li, X. Xing, X. Hou, T. Wang, J. Wang, and F. Xiao, "Determination of SARA fractions in asphalts by mid-infrared spectroscopy and multivariate calibration," *Measurement*, vol. 198, 2022, Art. no. 111361.

[29] S. Zhang et al., "Short wavelength infrared (SWIR) spectroscopy of phyllosilicate minerals from the Tonglushan Cu-Au-Fe deposit, Eastern China: New exploration indicators for concealed skarn orebodies," *Ore Geol. Rev.*, vol. 122, 2020, Art. no. 103516.

[30] S. Kumar, S. Arya, and K. Jain, "A SWIR-based vegetation index for change detection in land cover using multi-temporal landsat satellite dataset," *Int. J. Inf. Technol.*, vol. 14, no. 4, pp. 2035–2048, Sep. 2022.

[31] P. Nuradili et al., "UAV remote-sensing image semantic segmentation strategy based on thermal infrared and multispectral image features," *IEEE J. Miniaturization Air Space Syst.*, vol. 4, no. 3, pp. 311–319, Sep. 2023.

[32] J. Xu, K. Lu, and H. Wang, "Attention fusion network for multi-spectral semantic segmentation," *Pattern Recognit. Lett.*, vol. 146, pp. 179–184, 2021.

[33] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 115–134, 2019.

[34] A. Alem and S. Kumar, "Deep learning methods for land cover and land use classification in remote sensing: A review," in *Proc. 8th Int. Conf. Rel., Infocom Technol. Optim.*, 2020, pp. 903–908.

[35] M. S. Thasveen and S. Suresh, "Land–use and land–cover classification methods: A review," in *Proc. 4th Int. Conf. Microelectron., Signals Syst.*, 2021, pp. 1–6.

[36] M. Digra, R. Dhir, and N. Sharma, "Land use land cover classification of remote sensing images based on the deep learning approaches: A statistical analysis and review," *Arabian J. Geosci.*, vol. 15, no. 10, May 2022, Art. no. 1003.

[37] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021, Art. n71.

[38] M. Chamling and B. Bera, "Spatio-temporal patterns of land use/land cover change in the Bhutan–Bengal foothill region between 1987 and 2019: Study towards geospatial applications and policy making," *Earth Syst. Environ.*, vol. 4, no. 1, pp. 117–130, Mar. 2020.

[39] F. Thonfeld et al., "The impact of anthropogenic land use change on the protected areas of the kilombero catchment, Tanzania," *ISPRS J. Photogrammetry Remote Sens.*, vol. 168, pp. 41–55, 2020.

[40] Y. Zhou, X. Li, and Y. Liu, "Land use change and driving factors in rural China during the period 1995–2015," *Land Use Policy*, vol. 99, 2020, Art. no. 105048.

[41] D. C. Phan et al., "First comprehensive quantification of annual land use/cover from 1990 to 2020 across mainland Vietnam," *Sci. Rep.*, vol. 11, no. 1, May 2021, Art. no. 9979.

[42] C. J. Owers et al., "Operational continental-scale land cover mapping of Australia using the open data cube," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1715–1737, Oct. 2022.

[43] M. G. Tulbure, P. Hostert, T. Kuemmerle, and M. Broich, "Regional matters: On the usefulness of regional land-cover datasets in times of global change," *Remote Sens. Ecol. Conservation*, vol. 8, no. 3, pp. 272–283, Dec. 2021.

[44] T. Sarzana, A. Maltese, A. Capolupo, and E. Tarantino, "Post-processing of pixel and object-based land cover classifications of very high spatial resolution images," in *Proc. Int. Conf. Comput. Sci. Appl.*, 2020, pp. 797–812.

[45] T. Mollick, M. G. Azam, and S. Karim, "Geospatial-based machine learning techniques for land use and land cover mapping using a high-resolution unmanned aerial vehicle image," *Remote Sens. Appl.: Soc. Environ.*, vol. 29, 2023, Art. no. 100859.

[46] S. Shayeganpour, M. H. Tangestani, S. Homayouni, and R. K. Vincent, "Evaluating pixel-based vs. object-based image analysis approaches for lithological discrimination using VNIR data of WorldView-3," *Front. Earth Sci.*, vol. 15, no. 1, pp. 38–53, Mar. 2021.

[47] C. Zhang et al., "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, 2019.

[48] C. Liu, Y. Cao, C. Yang, Y. Zhou, and M. Ai, "Pattern identification and analysis for the traditional village using low altitude UAV-borne remote sensing: Multifeatured geospatial data to support rural landscape investigation, documentation and management," *J. Cultural Heritage*, vol. 44, pp. 185–195, 2020.

[49] H. Huang, Y. Lan, A. Yang, Y. Zhang, S. Wen, and J. Deng, "Deep learning versus object-based image analysis (OBIA) in weed mapping of UAV imagery," *Int. J. Remote Sens.*, vol. 41, no. 9, pp. 3446–3479, Jan. 2020.

[50] X. Zhang, P. Xiao, and X. Feng, "Object-specific optimization of hierarchical multiscale segmentations for high-spatial resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 308–321, 2020.

[51] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022.

[52] A. Sohail, N. A. Nawaz, A. A. Shah, S. Rasheed, S. Ilyas, and M. K. Ehsan, "A systematic literature review on machine learning and deep learning methods for semantic segmentation," *IEEE Access*, vol. 10, pp. 134557–134570, 2022.

[53] I. R. I. Haque and J. Neubert, "Deep learning approaches to biomedical image segmentation," *Inf. Med. Unlocked*, vol. 18, 2020, Art. no. 100297.

[54] D. J. Ho et al., "Deep multi-magnification networks for multi-class breast cancer image segmentation," *Comput. Med. Imag. Graph.*, vol. 88, 2021, Art. no. 101866.

[55] L. Ramos, E. Casas, E. Bendek, C. Romero, and F. Rivas-Echeverría, "Computer vision for wildfire detection: A critical brief review," *Multimedia Tools Appl.*, pp. 1–44, 2024, doi: 10.1007/s11042-024-18685-z.

[56] L. M. Tassis, J. E. T. de Souza, and R. A. Krohling, "A deep learning approach combining instance and semantic segmentation to identify diseases and pests of coffee leaves from in-field images," *Comput. Electron. Agriculture*, vol. 186, 2021, Art. no. 106191.

[57] S. Lan et al., "DiscoBox: Weakly supervised instance segmentation and semantic correspondence from box supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3386–3396.

[58] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, Jul. 2021.

[59] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: State of the art," *Int. J. Multimedia Inf. Retrieval*, vol. 9, no. 3, pp. 171–189, Jul. 2020.

[60] W. Gu, S. Bai, and L. Kong, "A review on 2D instance segmentation based on deep neural networks," *Image Vis. Comput.*, vol. 120, 2022, Art. no. 104401.

[61] D. R. Loh, W. X. Yong, J. Yapeter, K. Subburaj, and R. Chandramohanadas, "A deep learning approach to the screening of malaria infection: Automated and rapid cell counting, object detection and instance segmentation using mask R-CNN," *Comput. Med. Imag. Graph.*, vol. 88, 2021, Art. no. 101845.

[62] Z. Luo, W. Yang, Y. Yuan, R. Gou, and X. Li, "Semantic segmentation of agricultural images: A survey," *Inf. Process. Agriculture*, vol. 11, pp. 172–186, 2024.

[63] N. Kheradmandi and V. Mehranfar, "A critical review and comparative study on image segmentation-based techniques for pavement crack detection," *Construction Building Mater.*, vol. 321, 2022, Art. no. 126162.

[64] I. Qureshi et al., "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Inf. Fusion*, vol. 90, pp. 316–352, 2023.

[65] W. Zhou, X. Du, and S. Wang, "Techniques for image segmentation based on edge detection," in *Proc. IEEE Int. Conf. Comput. Sci., Electron. Inf. Eng. Intell. Control Technol.*, 2021, pp. 400–403.

[66] M. A. F. Malbog, L. L. Lacatan, R. M. Dellosa, Y. D. Austria, and C. F. Cunanan, "Edge detection comparison of hybrid feature extraction for combustible fire segmentation: A Canny vs Sobel performance analysis," in *Proc. 11th IEEE Control Syst. Graduate Res. Colloq.*, 2020, pp. 318–322.

[67] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[68] R. M. d. S. Pinto et al., "Land degradation mapping in the MATOPIBA region (Brazil) using remote sensing data and decision-tree analysis," *Sci. Total Environ.*, vol. 782, 2021, Art. no. 146900.

[69] F. Zhang and X. Yang, "Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection," *Remote Sens. Environ.*, vol. 251, 2020, Art. no. 112105.

[70] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 140–152, 2020.

[71] X. Yang, S. Zhang, J. Liu, Q. Gao, S. Dong, and C. Zhou, "Deep learning for smart fish farming: Applications, opportunities and challenges," *Rev. Aquaculture*, vol. 13, no. 1, pp. 66–90, Jun. 2020.

[72] A. Shoeibi et al., "Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review," *Comput. Biol. Med.*, vol. 136, 2021, Art. no. 104697.

[73] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, 2021, Art. no. 114417.

[74] L. Alzubaidi et al., "A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications," *J. Big Data*, vol. 10, no. 1, 2023, Art. no. 46.

[75] P. Jadhav, G. Rajguru, D. Datta, and S. Mukhopadhyay, "Automatic sleep stage classification using time–frequency images of CWT and transfer learning using convolution neural network," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 494–504, 2020.

[76] J. Iqbal, M. Vogt, and J. Bajorath, "Activity landscape image analysis using convolutional neural networks," *J. Cheminform.*, vol. 12, no. 1, 2020, Art. no. 34.

[77] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020.

[78] A. Girdhar, H. Kapur, and V. Kumar, "Classification of white blood cell using convolution neural network," *Biomed. Signal Process. Control*, vol. 71, 2022, Art. no. 103156.

[79] L. Zhang, X. Chen, J. Zhang, R. Dong, and K. Ma, "Contrastive deep super vision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–19.

[80] X. Cheng et al., "Enhanced contextual representation with deep neural networks for land cover classification based on remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 107, 2022, Art. no. 102706.

[81] D. R. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: A survey," *Evol. Intell.*, vol. 15, no. 1, pp. 1–22, Jan. 2021.

[82] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[83] Y. Li et al., "Fully convolutional networks for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 214–223.

[84] A. Abdollahi, B. Pradhan, and A. M. Alamri, "An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images," *Geocarto Int.*, vol. 37, no. 12, pp. 3355–3370, Dec. 2020.

[85] D. Zhang, J. Zhao, J. Chen, Y. Zhou, B. Shi, and R. Yao, "Edge-aware and spectral–spatial information aggregation network for multispectral image semantic segmentation," *Eng. Appl. Artif. Intell.*, vol. 114, 2022, Art. no. 105070.

[86] A. M. Ali et al., "Crop yield prediction using multi sensors remote sensing (review article)," *Egyptian J. Remote Sens. Space Sci.*, vol. 25, no. 3, pp. 711–716, 2022.

[87] V. Pejović, E. Georgitzikis, I. Lieberman, P. E. Malinowski, P. Heremans, and D. Cheyns, "Photodetectors based on lead sulfide quantum dot and organic absorbers for multispectral sensing in the visible to short-wave infrared range," *Adv. Funct. Mater.*, vol. 32, no. 28, Apr. 2022, Art. no. 2201424.

[88] S. M. Jameel, A. R. Gilal, S. S. H. Rizvi, M. Rehman, and M. A. Hashmani, "Practical implications and challenges of multispectral image analysis," in *Proc. 3rd Int. Conf. Comput., Math. Eng. Technol.*, 2020, pp. 1–5.

[89] M. Novák et al., "Multisensor UAV system for the forest monitoring," in *Proc. 10th Int. Conf. Adv. Comput. Inf. Technol.*, 2020, pp. 293–296.

[90] W. A. León-Rueda, C. León, S. G. Caro, and J. G. Ramírez-Gil, "Identification of diseases and physiological disorders in potato via multispectral drone imagery using machine learning tools," *Trop. Plant Pathol.*, vol. 47, no. 1, pp. 152–167, Sep. 2021.

[91] M. S. Navin and L. Agilandeeswari, "Multispectral and hyperspectral images based land use / land cover change prediction analysis: An extensive review," *Multimedia Tools Appl.*, vol. 79, nos. 39/40, pp. 29751–29774, Aug. 2020.

[92] F. Oriani, M. F. McCabe, and G. Mariethoz, "Downscaling multispectral satellite images without colocated high-resolution data: A stochastic approach based on training images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3209–3225, Apr. 2021.

[93] J. Yang, L. Xiao, Y.-Q. Zhao, and J. C.-W. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5508817.

[94] U. A. Bhatti et al., "Local similarity-based spatial–spectral fusion hyperspectral image classification with deep CNN and Gabor filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514215.

[95] I. F. Barton, M. J. Gabriel, J. Lyons-Baral, M. D. Barton, L. Duplessis, and C. Roberts, "Extending geometallurgy to the mine scale with hyperspectral imaging: A pilot study using drone- and ground-based scanning," *Mining, Metall. Exploration*, vol. 38, no. 2, pp. 799–818, Feb. 2021.

[96] Y. Tan, L. Lu, L. Bruzzone, R. Guan, Z. Chang, and C. Yang, "Hyperspectral band selection for lithologic discrimination and geological mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 471–486, 2020.

[97] L. Yao, M. Xu, Y. Liu, R. Niu, X. Wu, and Y. Song, "Estimating of heavy metal concentration in agricultural soils from hyperspectral satellite sensor imagery: Considering the sources and migration pathways of pollutants," *Ecol. Indicators*, vol. 158, 2024, Art. no. 111416.

[98] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

[99] S. Mei, Y. Geng, J. Hou, and Q. Du, "Learning hyperspectral images from RGB images via a coarse-to-fine CNN," *Sci. China Inf. Sci.*, vol. 65, no. 5, Sep. 2021, Art. no. 152102.

[100] K. M. Vignesh and Y. Kiran, "Comparative analysis of mineral mapping for hyperspectral and multispectral imagery," *Arabian J. Geosci.*, vol. 13, no. 4, Feb. 2020, Art. no. 160.

[101] F. Wang et al., "Combining spectral and textural information in UAV hyperspectral images to estimate rice grain yield," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, 2021, Art. no. 102397.

[102] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao, "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing," *J. Forestry Res.*, vol. 32, no. 1, pp. 1–6, May 2020.

[103] Q. Wang, Á. Moreno-Martínez, J. Muñoz-Marí, M. Campos-Taberner, and G. Camps-Valls, "Estimation of vegetation traits with kernel NDVI," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 408–417, 2023.

[104] G. L. Spadoni, A. Cavalli, L. Congedo, and M. Munafò, "Analysis of normalized difference vegetation index (NDVI) multi-temporal series for the production of forest cartography," *Remote Sens. Appl.: Soc. Environ.*, vol. 20, 2020, Art. no. 100419.

[105] A. K. Taloor, D. S. Manhas, and G. C. Kothyari, "Retrieval of land surface temperature, normalized difference moisture index, normalized difference water index of the Ravi Basin using Landsat data," *Appl. Comput. Geosci.*, vol. 9, 2021, Art. no. 100051.

[106] H. Chafik, M. Berrada, A. Legdou, A. Amine, and S. Lahssini, "Exploitation of spectral indices NDVI, NDWI & SAVI in random forest classifier model for mapping weak rosemary cover: Application on Gourrama region, Morocco," in *Proc. IEEE Int. Conf. Moroccan Geomatics*, 2020, pp. 1–6.

[107] Y. Wang, Z. Li, C. Zeng, G.-S. Xia, and H. Shen, "An urban water extraction method combining deep learning and Google Earth engine," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 769–782, 2020.

[108] G. Yang, K. Huang, W. Sun, X. Meng, D. Mao, and Y. Ge, "Enhanced mangrove vegetation index based on hyperspectral images for mapping mangrove," *ISPRS J. Photogrammetry Remote Sens.*, vol. 189, pp. 236–254, 2022.

[109] F. Mirchooli, M. Kiani-Harchegani, A. K. Darvishan, S. Falahatkar, and S. H. Sadeghi, "Spatial distribution dependency of soil organic carbon content to important environmental variables," *Ecol. Indicators*, vol. 116, 2020, Art. no. 106473.

[110] C. Alexander, "Normalised difference spectral indices and urban land cover as indicators of land surface temperature (LST)," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 86, 2020, Art. no. 102013.

[111] P. Rhyma, K. Norizah, O. Hamdan, I. Faridah-Hanum, and A. Zulfa, "Integration of normalised different vegetation index and soil-adjusted vegetation index for mangrove vegetation delineation," *Remote Sens. Appl.: Soc. Environ.*, vol. 17, 2020, Art. no. 100280.

[112] E. M. Domínguez, D. Small, and D. Henke, "Deriving digital surface models from geocoded SAR images and back-projection tomography," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4339–4351, 2021.

[113] F. Schiefer et al., "Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 170, pp. 205–215, 2020.

[114] N. M. Enwright et al., "Developing bare-earth digital elevation models from structure-from-motion data on barrier islands," *ISPRS J. Photogrammetry Remote Sens.*, vol. 180, pp. 269–282, 2021.

[115] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, 2020, Art. no. 111322.

[116] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

[117] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.

[118] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1524–1532.

[119] T. L. Giang, K. B. Dang, Q. T. Le, V. G. Nguyen, S. S. Tong, and V.-M. Pham, "U-Net convolutional networks for mining land cover classification based on high-resolution UAV imagery," *IEEE Access*, vol. 8, pp. 186257–186273, 2020.

[120] N. Pahlevan et al., "ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and sentinel-2 over lakes, rivers, and coastal waters," *Remote Sens. Environ.*, vol. 258, 2021, Art. no. 112366.

[121] S. Peyghambari and Y. Zhang, "Hyperspectral remote sensing in lithological mapping, mineral exploration, and environmental geology: An updated review," *J. Appl. Remote Sens.*, vol. 15, no. 3, 2021, Art. no. 031501.

[122] X. Che, H. K. Zhang, and J. Liu, "Making landsat 5, 7 and 8 reflectance consistent using MODIS nadir-BRDF adjusted reflectance as reference," *Remote Sens. Environ.*, vol. 262, 2021, Art. no. 112517.

[123] J. Pancorbo, B. Lamb, M. Quemada, W. Hively, I. Gonzalez-Fernandez, and I. Molina, "Sentinel-2 and WorldView-3 atmospheric correction and signal normalization based on ground-truth spectroradiometric measurements," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 166–180, 2021.

[124] N. Sanchiz-Viel, E. Bretagne, E. M. Mouaddib, and P. Dassonvalle, "Radiometric correction of laser scanning intensity data applied for terrestrial laser scanning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 172, pp. 1–16, 2021.

[125] L. Li, G. Zhang, Y. Jiang, and X. Shen, "An improved on-orbit relative radiometric calibration method for agile high-resolution optical remote-sensing satellites with sensor geometric distortion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606715.

[126] A. A. Darem, A. A. Alhashmi, A. M. Almadani, A. K. Alanazi, and G. A. Sutantra, "Development of a map for land use and land cover classification of the northern border region using remote sensing and GIS," *Egyptian J. Remote Sens. Space Sci.*, vol. 26, no. 2, pp. 341–350, 2023.

[127] X. Zhou et al., "Radiometric calibration of a large-array commodity CMOS multispectral camera for UAV-borne remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102968.

[128] P. Dayal, P. Goel, C. Gupta, and T. K. Patra, "Change detection using relative radiometric correction on air field satellite imagery," in *Proc. 8th Int. Conf. Rel., Infocom Technol. Optim.*, 2020, pp. 753–757.

[129] C. Gudavalli, M. Goebel, T. Nanjundaswamy, L. Nataraj, S. Chandrasekaran, and B. S. Manjunath, "Resampling estimation based RPC metadata verification in satellite imagery," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 3212–3221, 2023.

[130] T. Wang et al., "Large-scale orthorectification of GF-3 SAR images without ground control points for China's land area," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5221617.

[131] A. Ulvi, "The effect of the distribution and numbers of ground control points on the precision of producing orthophoto maps with an unmanned aerial vehicle," *J. Asian Archit. Building Eng.*, vol. 20, no. 6, pp. 806–817, Sep. 2021.

[132] R. Zhou et al., "A large-batch orthorectification generation method based on adaptive GPU thread parameters and parallel calculation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4638–4648, 2023.

[133] E. Zhang, L. Liu, L. Huang, and K. S. Ng, "An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery," *Remote Sens. Environ.*, vol. 254, 2021, Art. no. 112265.

[134] K. Malik, C. Robertson, D. Braun, and C. Greig, "U-Net convolutional neural network models for detecting and quantifying placer mining disturbances at watershed scales," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, 2021, Art. no. 102510.

[135] X. Lv et al., "Pruning for image segmentation: Improving computational efficiency for large-scale remote sensing applications," *ISPRS J. Photogrammetry Remote Sens.*, vol. 202, pp. 13–29, 2023.

[136] Q. Zhu et al., "Land-use/land-cover change detection based on a siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 63–78, 2022.

[137] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[138] C. Zhang, P. Yue, D. Tapete, B. Shangguan, M. Wang, and Z. Wu, "A multi-level context-guided classification method with object-based convolutional neural network for land cover classification using very high resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 88, 2020, Art. no. 102086.

[139] Q. Hu, Y. Wu, and Y. Li, "Semi-supervised semantic labeling of remote sensing images with improved image-level selection retraining," *Alexandria Eng. J.*, vol. 94, pp. 235–247, 2024.

[140] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, 2021.

[141] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. IEEE 23rd Int. Conf. Inf. Fusion*, 2020, pp. 1–6.

[142] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, 2021, Art. no. 104042.

[143] M. Brenner, N. H. Reyes, T. Susnjak, and A. L. C. Barczak, "RGB-D and thermal sensor fusion: A systematic literature review," *IEEE Access*, vol. 11, pp. 82410–82442, 2023.

[144] R. Soroush and Y. Baleghi, "NIR/RGB image fusion for scene classification using deep neural networks," *Vis. Comput.*, vol. 39, pp. 2725–2739, May 2022.

[145] F. Moradi, F. D. Javan, and F. Samadzadegan, "Potential evaluation of visible-thermal UAV image fusion for individual tree detection based on convolutional neural network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, 2022, Art. no. 103011.

[146] J. Li, X. Cai, and J. Qi, "AMFNet: An attention-based multi-level feature fusion network for ground objects extraction from mining area's UAV-based RGB images and digital surface model," *J. Appl. Remote Sens.*, vol. 15, no. 3, 2021, Art. no. 036506.

[147] J. Zheng, Y. Feng, C. Bai, and J. Zhang, "Hyperspectral image classification using mixed convolutions and covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 522–534, Jan. 2021.

[148] H. Kaur, D. Koundal, and V. Kadyan, "Image fusion techniques: A survey," *Arch. Comput. Methods Eng.*, vol. 28, no. 7, pp. 4425–4447, Jan. 2021.

[149] S. Aymaz, C. Kose, and S. Aymaz, "Multi-focus image fusion for different datasets with super-resolution using gradient-based new fusion rule," *Multimedia Tools Appl.*, vol. 79, no. 19/20, pp. 13311–13350, Jan. 2020.

[150] S. P. Yadav and S. Yadav, "Image fusion using hybrid methods in multimodality medical images," *Med. Biol. Eng. Comput.*, vol. 58, no. 4, pp. 669–687, Jan. 2020.

[151] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: A review," *Complex Intell. Syst.*, vol. 8, pp. 2663–2693, Jan. 2022.

[152] R. Zhai, J. Zeng, and Z. Ge, "Structured principal component analysis model with variable correlation constraint," *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 2, pp. 558–569, Mar. 2022.

[153] S. Bhat and D. Koundal, "Multi-focus image fusion using neutrosophic based wavelet transform," *Appl. Soft Comput.*, vol. 106, 2021, Art. no. 107307.

[154] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 238–262, 2021.

[155] S. T. Yekeen, A. Balogun, and K. B. W. Yusof, "A novel deep learning instance segmentation model for automated marine oil spill detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 190–200, 2020.

[156] J. Liu, D. Zhang, L. He, X. Yu, and W. Han, "MFAGNet: Multi-scale frequency attention gating network for land cover classification," *Int. J. Remote Sens.*, vol. 44, no. 21, pp. 6670–6697, Nov. 2023.

[157] A. Nanda, R. C. Barik, and S. Bakshi, "SSO-RBNN driven brain tumor classification with saliency-K-means segmentation technique," *Biomed. Signal Process. Control*, vol. 81, 2023, Art. no. 104356.

[158] Y. Xie, R. Chen, M. Yu, X. Rui, and X. Du, "Improvement and application of Unet network for avoiding the effect of urban dense high-rise buildings and other feature shadows on water body extraction," *Int. J. Remote Sens.*, vol. 44, no. 12, pp. 3861–3891, 2023.

[159] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[160] H. Yuan, Z. Zhang, X. Rong, D. Feng, S. Zhang, and S. Yang, "MPFFNet: LULC classification model for high-resolution remote sensing images with multi-path feature fusion," *Int. J. Remote Sens.*, vol. 44, no. 19, pp. 6089–6116, Oct. 2023.

[161] M. Vergara, L. Ramos, N. D. Rivera-Campoverde, and F. Rivas-Echeverría, "EngineFaultDB: A novel dataset for automotive engine fault classification and baseline results," *IEEE Access*, vol. 11, pp. 126155–126171, 2023.

[162] A. Perera, B. Turhan, A. Aleti, and M. Böhme, "On the impact of lower recall and precision in defect prediction for guiding search-based software testing," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, pp. 1–27, Apr. 2024.

[163] M. S. V. D. P. Jasti and K. S. Vani, "Land cover change detection based on LeNet-5 by using very-high-resolution remote sensing images," *IETE J. Res.*, pp. 1–13, 2023.

[164] Q. Ni et al., "A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images," *Eur. Radiol.*, vol. 30, no. 12, pp. 6517–6527, 2020.

[165] S. Jombo, E. Adam, and J. Odindi, "Classification of tree species in a heterogeneous urban environment using object-based ensemble analysis and world view-2 satellite imagery," *Appl. Geomatics*, vol. 13, no. 3, pp. 373–387, Jan. 2021.

[166] G. O. Tetteh, M. Schwieder, S. Erasmi, C. Conrad, and A. Gocht, "Comparison of an optimised multiresolution segmentation approach with deep neural networks for delineating agricultural fields from sentinel-2 images," *PFG—J. Photogrammetry, Remote Sens. Geoinf. Sci.*, vol. 91, no. 4, pp. 295–312, Jun. 2023.

[167] A. B. Polat, O. Akcay, and F. B. Sanli, "Monitoring seasonal effects in vegetation areas with sentinel-1 SAR and sentinel-2 optic satellite images," *Arabian J. Geosci.*, vol. 15, no. 7, Mar. 2022, Art. no. 670.

[168] P. Nagarajan, L. Rajendran, N. D. Pillai, and G. Lakshmanan, “Comparison of machine learning algorithms for mangrove species identification in Malad creek, Mumbai using worldview-2 and Google Earth images,” *J. Coastal Conservation*, vol. 26, no. 5, Sep. 2022, Art. no. 44.

[169] P. Esmaeili, M. Vazifedoust, M. Rahmani, and H. Pakdel, “A simple rule-based algorithm in Google Earth Engine for operational discrimination of rice paddies in Sefidroud Irrigation Network,” *Arabian J. Geosci.*, vol. 16, no. 12, Nov. 2023, Art. no. 649.

[170] Z. Zhao et al., “Comparison of three machine learning algorithms using Google Earth Engine for land use land cover classification,” *Rangeland Ecol. Manage.*, vol. 92, pp. 129–137, 2024.

[171] S. Qamar, H. Jin, R. Zheng, P. Ahmad, and M. Usama, “A variant form of 3D-UNet for infant brain segmentation,” *Future Gener. Comput. Syst.*, vol. 108, pp. 613–623, 2020.

[172] X. Zheng and T. Chen, “Segmentation of high spatial resolution remote sensing image based on U-Net convolutional networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 2571–2574.

[173] X. Zheng and T. Chen, “High spatial resolution remote sensing image segmentation based on the multiclassification model and the binary classification model,” *Neural Comput. Appl.*, vol. 35, no. 5, pp. 3597–3604, Jan. 2021.

[174] K. Chaurasia, R. Nandy, O. Pawar, R. R. Singh, and M. Ahire, “Semantic segmentation of high-resolution satellite images using deep learning,” *Earth Sci. Informat.*, vol. 14, no. 4, pp. 2161–2170, Aug. 2021.

[175] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[176] J. Perez-Guerra, V. Herrera-Ruiz, J. C. Gonzalez-Velez, J. D. Martinez-Vargas, and M. C. Torres-Madronero, “Land cover classification using remote sensing and supervised convolutional neural networks,” in *Advances in Computing*, M. Tabares, P. Vallejo, B. Suarez, M. Suarez, O. Ruiz, and J. Aguilar, Eds., Cham, Switzerland: Springer, 2024, pp. 13–24.

[177] D. Sathyanarayanan et al., “A multiclass deep learning approach for LULC classification of multispectral satellite images,” in *Proc. IEEE India Geosci. Remote Sens. Symp.*, 2020, pp. 102–105.

[178] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[179] R. Sivagami, J. Srihari, and K. S. Ravichandran, “Analysis of encoder-decoder based deep learning architectures for semantic segmentation in remote sensing images,” in *Intelligent Systems Design and Applications*, A. Abraham, A. K. Cherukuri, P. Melin, and N. Gandhi, Eds., Cham, Switzerland: Springer, 2020, pp. 332–341.

[180] W. Xu, Y.-L. Fu, and D. Zhu, “Resnet and its application to medical image processing: Research progress and challenges,” *Comput. Methods Programs Biomed.*, vol. 240, 2023, Art. no. 107660.

[181] L. Borawar and R. Kaur, “ResNet: Solving vanishing gradient in deep networks,” in *Proc. Int. Conf. Recent Trends Comput.*, 2023, pp. 235–247.

[182] N. Saxena, K. B. N, and B. Raman, “Semantic segmentation of multispectral images using Res-Seg-net model,” in *Proc. IEEE 14th Int. Conf. Semantic Comput.*, 2020, pp. 154–157.

[183] Y. Fan, X. Ding, J. Wu, J. Ge, and Y. Li, “High spatial-resolution classification of urban surfaces using a deep learning method,” *Building Environ.*, vol. 200, 2021, Art. no. 107949.

[184] N. J. Singh and K. Nongmeikapam, “Semantic segmentation of satellite images using deep-Unet,” *Arabian J. Sci. Eng.*, vol. 48, no. 2, pp. 1193–1205, Mar. 2022.

[185] M. A. Aliyu, S. Boukari, A. M. Gamsha, M. L. Abdurrahman, and A. Y. Gital, “Toward a better model for the semantic segmentation of remote sensing imagery,” in *Proc. 3rd Int. Conf. Artif. Intell.: Adv. Appl.*, 2023, pp. 407–415.

[186] Y. Ma, B. M. Narayanaswamy, H. Lin, and H. Ding, “Temporal-contextual recommendation in real-time,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2291–2299.

[187] J. Duan, P.-F. Zhang, R. Qiu, and Z. Huang, “Long short-term enhanced memory for sequential recommendation,” *World Wide Web*, vol. 26, pp. 561–583, May 2022.

[188] Z. Niu, G. Zhong, and H. Yu, “A review on the attention mechanism of deep learning,” *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[189] W. He, Y. Wu, and X. Li, “Attention mechanism for neural machine translation: A survey,” in *Proc. IEEE 5th Inf. Technol., Netw., Electron. Autom. Control Conf.*, 2021, pp. 1485–1489.

[190] C. Yang, J. Hou, and Y. Wang, “Extraction of land covers from remote sensing images based on a deep learning model of NDVI-RSU-Net,” *Arabian J. Geosci.*, vol. 14, no. 20, Sep. 2021, Art. no. 2073.

[191] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[192] Z. Nong, X. Su, Y. Liu, Z. Zhan, and Q. Yuan, “Boundary-aware dual-stream network for VHR remote sensing images semantic segmentation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5260–5268, 2021.

[193] J. Huang, X. Zhang, Y. Sun, and Q. Xin, “Attention-guided label refinement network for semantic segmentation of very high resolution aerial orthoimages,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4490–4503, 2021.

[194] Z. Liang and X. Wang, “Semantic segmentation network with band-location adaptive selection mechanism for multispectral remote sensing images,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 3488–3491.

[195] L. Ding, H. Tang, and L. Bruzzone, “LANet: Local attention embedding to improve the semantic segmentation of remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[196] K. Zhang and H. Yang, “Semi-supervised multi-spectral land cover classification with multi-attention and adaptive kernel,” in *2020 IEEE Int. Conf. Image Process.*, 2020, pp. 1881–1885.

[197] J. Chen, H. Wang, Y. Guo, G. Sun, Y. Zhang, and M. Deng, “Strengthen the feature distinguishability of geo-object details in the semantic segmentation of high-resolution remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2327–2340, 2021.

[198] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, “Hybrid multiple attention network for semantic segmentation in aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603018.

[199] W. Wang and X. Wang, “A novel semantic segmentation method for high-resolution remote sensing images based on visual attention network,” in *Proc. Int. Conf. Image Graph.*, 2023, pp. 42–53.

[200] R. Li et al., “Multiattention network for semantic segmentation of fine-resolution remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.

[201] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021, *arXiv:2010.11929v2*.

[202] Z. Chen et al., “DPT: Deformable patch-based transformer for visual recognition,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2899–2907.

[203] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, “Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408820.

[204] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, “A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6506105.

[205] X. Zhou, L. Zhou, S. Gong, S. Zhong, W. Yan, and Y. Huang, “Swin transformer embedding dual-stream for semantic segmentation of remote sensing imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 175–189, 2024.

[206] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[207] X. Li, Y. Cheng, Y. Fang, H. Liang, and S. Xu, “2DSegFormer: 2-D transformer model for semantic segmentation on aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4709413.

[208] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[209] S. Du and M. Liu, “Class-guidance network based on the pyramid vision transformer for efficient semantic segmentation of high-resolution remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5578–5589, 2023.

[210] W. Wang et al., “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *2021 IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 548–558.

[211] J. Chen, J. Yi, A. Chen, and H. Lin, “SRCBTFusion-Net: An efficient fusion architecture via stacked residual convolution blocks and transformer for remote sensing image semantic segmentation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411716.

[212] L. Gao et al., "STransFuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10990–11003, 2021.

[213] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[214] Y. Liu, Y. Zhang, Y. Wang, and S. Mei, "Rethinking transformers for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617515.

[215] Y. Chen et al., "Hybrid attention fusion embedded in transformer for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4421–4435, 2024.

[216] A. Qayyum, I. Ahmad, W. Mumtaz, M. O. Alassafi, R. Alghamdi, and M. Mazher, "Automatic segmentation using a hybrid dense network integrated with an 3D-atrous spatial pyramid pooling module for computed tomography (CT) imaging," *IEEE Access*, vol. 8, pp. 169794–169803, 2020.

[217] Z. Li, H. Zhu, and M. Huang, "A deep learning-based fine crack segmentation network on full-scale steel bridge images with complicated backgrounds," *IEEE Access*, vol. 9, pp. 114989–114997, 2021.

[218] Y. Huang, Q. Wang, W. Jia, Y. Lu, Y. Li, and X. He, "See more than once: Kernel-sharing atrous convolution for semantic segmentation," *Neurocomputing*, vol. 443, pp. 26–34, 2021.

[219] B. Hou et al., "Panchromatic image land cover classification via DCNN with updating iteration strategy," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1472–1475.

[220] B. Zhang, Y. Wan, Y. Zhang, and Y. Li, "JSH-Net: Joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery," *Int. J. Remote Sens.*, vol. 43, no. 17, pp. 6307–6332, Sep. 2022.

[221] Y. Wang, W. Ding, R. Zhang, and H. Li, "Boundary-aware multitask learning for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 951–963, 2021.

[222] J. Jin et al., "DASFNet: Dense-attention-similarity-fusion network for scene classification of dual-modal remote-sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 115, 2022, Art. no. 103087.

[223] Y. Li, L. Wang, L. Zhang, H. Chen, S. Wang, and X. Wang, "HrreNet: Semantic segmentation network for moderate and high-resolution satellite images," *Int. J. Remote Sens.*, vol. 43, no. 11, pp. 4065–4086, Jun. 2022.

[224] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[225] Z. Zheng, X. Zhang, P. Xiao, and Z. Li, "Integrating gate and attention modules for high-resolution image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4530–4546, 2021.

[226] W. Zhou, X. Fan, W. Yan, S. Shan, Q. Jiang, and J.-N. Hwang, "Graph attention guidance network with knowledge distillation for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4506015.

[227] J. Jin, W. Zhou, R. Yang, L. Ye, and L. Yu, "Edge detection guide network for semantic segmentation of remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5000505.

[228] Y. Ni, J. Liu, J. Cui, Y. Yang, and X. Wang, "Edge guidance network for semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9382–9395, 2023.

[229] X. Sun, Y. Qian, R. Cao, P. Tuerxun, and Z. Hu, "BGFNet: Semantic segmentation network based on boundary guidance," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2500305.

[230] H. Han, Z. Feng, W. Du, S. Guo, P. Wang, and T. Xu, "Remote sensing image classification based on multi-spectral cross-sensor super-resolution combined with texture features: A case study in the Liaohe planting area," *IEEE Access*, vol. 12, pp. 16830–16843, 2024.

[231] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[232] R. Alshehhi and P. R. Marpu, "Extraction of urban multi-class from high-resolution images using pyramid generative adversarial networks," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, 2021, Art. no. 102379.

[233] C. Wang, B. Chen, Z. Zou, and Z. Shi, "Remote sensing image synthesis via semantic embedding generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702811.

[234] C. Shi, L. Fang, Z. Lv, and H. Shen, "Improved generative adversarial networks for VHR remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8001805.

[235] X. Pan, J. Zhao, and J. Xu, "Conditional generative adversarial network-based training sample set improvement model for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7854–7870, Sep. 2021.

[236] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[237] B. Sui, T. Jiang, Z. Zhang, and X. Pan, "ECGAN: An improved conditional generative adversarial network with edge detection to augment limited training data for the classification of remote sensing images with high spatial resolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1311–1325, 2021.

[238] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*.

[239] M. E. Paoletti, J. M. Haut, P. Ghamisi, N. Yokoya, J. Plaza, and A. Plaza, "U-IMG2DSM: Unpaired simulation of digital surface models with generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 7, pp. 1288–1292, Jul. 2021.

[240] C. J. Costa, S. Tiwari, K. Bhagat, A. Verlekar, K. M. C. Kumar, and S. Aswale, "Three-dimensional reconstruction of satellite images using generative adversarial networks," in *Proc. Int. Conf. Technol. Adv. Innovations*, 2021, pp. 121–126.

[241] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[242] B. Kolbeinsson and K. Mikolajczyk, "Multi-class segmentation from aerial views using recursive noise diffusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 8424–8434.

[243] F. Xu, Z. Shang, Q. Wu, X. Zhang, Z. Lin, and S. Shao, "MUFNet: Toward semantic segmentation of multi-spectral remote sensing images," in *Proc. 4th Artif. Intell. Cloud Comput. Conf.*, 2022, pp. 39–46.

[244] Y. Wang, Q. Li, Y. Liu, and W. Wang, "A general dual-branch framework for land cover mapping models with multispectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2503305.

[245] J. Jiang, X. Feng, and H. Huang, "Semantic segmentation of remote sensing images based on dual-channel attention mechanism," *IET Image Process.*, vol. 18, pp. 2346–2356, 2024.

[246] W.-K. Baek, M.-J. Lee, and H.-S. Jung, "Land cover classification from RGB and NIR satellite images using modified U-Net model," *IEEE Access*, vol. 12, pp. 69445–69455, 2024.

[247] Y. Cao, Y. Shi, Y. Liu, C. Huo, S. Xiang, and C. Pan, "Dual stream fusion network for multi-spectral high resolution remote sensing image segmentation," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2021, pp. 537–547.

[248] Y.-C. Li, H.-C. Li, W.-S. Hu, and H.-L. Yu, "DSPCANet: Dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8552–8565, 2021.

[249] Q. Weng, H. Chen, H. Chen, W. Guo, and Z. Mao, "A multisensor data fusion model for semantic segmentation in aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6511905.

[250] W. Zhou, J. Jin, J. Lei, and L. Yu, "CIMFNet: Cross-layer interaction and multiscale fusion network for semantic segmentation of high-resolution remote sensing images," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 666–676, Jun. 2022.

[251] W. Zhou, Y. Li, J. Huan, Y. Liu, and Q. Jiang, "MSTNet-KD: Multilevel transfer networks using knowledge distillation for the dense prediction of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4504612.

[252] C. García, O. Mora, F. Pérez-Aragüés, and J. Vitrià, "CatLC: Catalonia multiresolution land cover dataset," *Sci. Data*, vol. 9, no. 1, Sep. 2022, Art. no. 554.

[253] X.-Y. Tong, G.-S. Xia, and X. X. Zhu, "Enabling country-scale land cover mapping with meter-resolution satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 178–196, 2023.

[254] L. Abady, M. Barni, A. Garzelli, and B. Tondi, "GAN generation of synthetic multispectral satellite images," in *Proc. SPIE*, vol. 11533, 2020, Art. no. 115330L.

[255] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Synthesizing optical and SAR imagery from land cover maps and auxiliary raster data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 4701312.

[256] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.

Leo Thomas Ramos (Member, IEEE) received the information technology engineering degree (*magna cum laude*) with a focus on artificial intelligence from Yachay Tech University, Urcuqui, Ecuador, in 2023. He is currently working toward the Ph.D. degree in computer science with the Computer Vision Center, Universitat Autónoma de Barcelona, Barcelona, Spain.

He is currently a Senior Research Engineer with Zeus Intelligent Solutions, Houston, TX, USA, where he is involved in the research and development of deep learning models for industry applications. In addition, he is an Applied Research Engineer with Kauel Inc., Menlo Park, Silicon Valley, CA, USA, contributing to the development of intelligent systems for the energy sector and scientific research. In line with his research interests, he collaborates with renowned researchers from Venezuela, Spain, France, U.K., the UAE, and the USA, including collaborations with NASA scientists. His research interests include deep learning, computer vision, natural language processing, and remote sensing.

Angel D. Sappa (Senior Member, IEEE) received the electromechanical engineering degree from the National University of La Pampa, General Pico, Argentina, in 1995, and the Ph.D. degree in industrial engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 1999.

In 2003, after holding research positions in France, the U.K., and Greece, he joined the Computer Vision Center, Barcelona, where he is currently a Senior Scientist. Since 2016, he has been a Full Professor with ESPOL Polytechnic University, Guayaquil, Ecuador, where he leads the computer vision team with the CIDIS Research Center. He is also the Director of the Electrical Engineering Ph.D. Program with ESPOL Polytechnic University.