1-29-2022

A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability

Rongjun Qin
The Ohio State University

Tao Liu
Michigan Technological University, taoliu@mtu.edu

Recommended Citation

Qin, R., & Liu, T. (2022). A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability. *Remote Sensing, 14*(3).
http://doi.org/10.3390/rs14030646
Retrieved from: https://digitalcommons.mtu.edu/michigantech-p/15726

Follow this and additional works at: https://digitalcommons.mtu.edu/michigantech-p

 Part of the Forest Sciences Commons

remote sensing

MDPI

Review

A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability

Rongjun Qin [1,2,3,4] and Tao Liu [5,6,*]

[1] Geospatial Data Analytics Lab, The Ohio State University, 218B Bolz Hall, 2036 Neil Avenue, Columbus, OH 43210, USA; qin.324@osu.edu
[2] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, 2070 Neil Avenue, Columbus, OH 43210, USA
[3] Department of Electrical and Computer Engineering, The Ohio State University, 205 Dreese Labs, 2015 Neil Avenue, Columbus, OH 43210, USA
[4] Translational Data Analytics Institute, The Ohio State University, Pomerene Hall, 1760 Neil Ave, Columbus, OH 43210, USA
[5] College of Forest Resources and Environmental Science, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA
[6] Ecosystem Science Center, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA
* Correspondence: taoliu@mtu.edu

Abstract: As an important application in remote sensing, landcover classification remains one of the most challenging tasks in very-high-resolution (VHR) image analysis. As the rapidly increasing number of Deep Learning (DL) based landcover methods and training strategies are claimed to be the state-of-the-art, the already fragmented technical landscape of landcover mapping methods has been further complicated. Although there exists a plethora of literature review work attempting to guide researchers in making an informed choice of landcover mapping methods, the articles either focus on the review of applications in a specific area or revolve around general deep learning models, which lack a systematic view of the ever advancing landcover mapping methods. In addition, issues related to training samples and model transferability have become more critical than ever in an era dominated by data-driven approaches, but these issues were addressed to a lesser extent in previous review articles regarding remote sensing classification. Therefore, in this paper, we present a systematic overview of existing methods by starting from learning methods and varying basic analysis units for landcover mapping tasks, to challenges and solutions on three aspects of scalability and transferability with a remote sensing classification focus including (1) sparsity and imbalance of data; (2) domain gaps across different geographical regions; and (3) multi-source and multi-view fusion. We discuss in detail each of these categorical methods and draw concluding remarks in these developments and recommend potential directions for the continued endeavor.

Keywords: very-high resolution; VHR; landcover classification; semantic segmentation; analysis unit; deep learning; transfer learning; data fusion; remote sensing

Citation: Qin, R.; Liu, T. A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability. *Remote Sens.* **2022**, *14*, 646. https://doi.org/10.3390/rs14030646

Academic Editor: Tais Grippa

Received: 7 December 2021
Accepted: 25 January 2022
Published: 29 January 2022

1. Introduction

Landcover mapping using remote sensing (RS) images has presented a consistent requirement for decades since the collection of the very first RS image. It greatly facilitates automated analysis of urban, suburban, and natural environments for applications such as urban expansion monitoring, change detection, crop prediction, forestation/deforestation, surveillance, anthropogenic activities, mining, etc. Generally, it is considered as a highly disparate problem and often appears to be application- and even location-dependent, as the learning systems need to accommodate the varying reference/training data with

quality and availability, the complexity of landcover classes, and the multi-source/multi-modal datasets. As of now, the accurate production of coarse resolution landcover classification maps at the global level (e.g., 30-m resolution) still follows a semi-automated approach and is often labor-intensive [1]. However, when it comes to high-resolution or very-high-resolution (VHR) images (resolution at the sub-meter level), there exist even greater challenges in classification tasks, due to the desired high-resolution output and level of uncertainty in predictions. Presently, with the collection of imageries from space-borne sensors/platforms such as WorldView constellations, IKONONS (decommissioned), GeoEye, Pleiades, Planet, etc., the volume of available VHR images has increased to an unprecedented level, and there exist a large body of approaches developed to address the classification of VHR data, from simple statistical learning based spectrum classification, spatial–spectral feature extraction, towards the recently popularized deep learning (DL) methods. Additionally, the available training data in varying multi-source forms (e.g., multi-view, multi-temporal, LiDAR data, or Synthetic Aperture Radar data), are becoming decisive when considering the scalability of landcover classifications.

Since the introduction of DL to the RS community, it has been widely adopted to solve a variety of RS tasks in classification, segmentation and object detection, and a few relevant review articles were published relevant to DL-based RS applications. For example, ref. [2] reviewed the DL models for road extraction during the time period from 2010 to 2019; ref. [3] discussed data sources, data preparation, training details and performance comparison for DL semantic segmentation models for satellite images in urban environments; refs. [4,5] reviewed DL applications in hyperspectral and multispectral images; [6,7] reviewed DL approaches to process 3D point cloud or RS data; ref. [8] reviewed various applications based on DL models including detection and segmentation of individual plants and vegetation classes; ref. [9] broadly reviewed applications of DL models on various RS tasks including image preprocessing, change detection and classification; ref. [10] discussed various DL models used for wetland mapping; ref. [11,12] reviewed 429 studies to investigate the impact of DL for earth observation applications through image segmentation and object detection; ref. [13] reviewed 416 papers in a meta-analysis and discussed the distribution of data sensors, DL models and application types in those studies; refs. [14,15] reviewed the deep learning applications for scene classification on aspects of the challenges, methods, benchmarks and opportunities.

While these literature surveys mainly focused on the diverse applications and the achievable accuracies of different models, few summarized works that inherently lead to scalable solutions to address the problem of training data sparsity of domain gaps. Moreover, existing reviews on RS classification, primarily introduced DL as a general approach or scene classification method, thus there remains a lack of a systematic taxonomy of DL methods specifically focusing on landcover mapping. In this paper, we aim to provide a comprehensive review of landcover classification methods from the perspective of the ever-growing data and scalability challenges. To begin, we first draw the connections between the DL-based approaches and traditional classifiers in terms of the analysis unit, to form an easily understandable taxonomy of DL approaches. Following these basics, we then outline its existing scalability challenges and summarize available solutions addressing them, and present our views on these potential solutions.

1.1. Scope and Organization of This Paper

In this paper, we aim to introduce existing works addressing issues (e.g., low quantity and quality of training samples, domain adaption, multimodal data fusion, etc.) related to landcover mapping using VHR images and outlook potential research directions. While our review focuses on works with applications to VHR RS data (data with Ground Sampling Distance of 2 m or less), we might introduce representative works on lower-resolution data when the relevant methods are of value to support VHR data for this review. Furthermore, although this review will largely involve the new opportunities and methods brought by DL, we maintain an equivalent emphasis on the relevant shallow classifiers as well,

especially when addressing framework-level approaches (semi-/weak- supervision) that are model agnostic. Given that there is a large body of literature related to general machine learning models and methodologies, this review only addresses approaches related to or adapted to RS image classifications.

The rest of this review is organized as follows: In Sections 1.2 and 1.3, we briefly review the existing issues of landcover mapping using VHR and the related efforts to address those issues; In Section 2, we provide a concise illustration of landcover classification paradigms using VHR from the perspective of the analysis unit to engage necessary contents and basics; In Section 3, we elaborate on existing approaches addressing the data challenges (briefly mentioned in Sections 1.2 and 1.3) for RS landcover classification. In Section 4, we conclude this review by discussing our findings and providing our outlooks on potential approaches to move existing practices forward.

1.2. Existing Challenges in the Landcover Classification with VHR Images

As the number of available VHR images, annotated data, and complexity of learning models continue to grow, there have been many landcover classification studies focusing on a diverse set of applications. Among these studies, the major challenges presented in RS image landcover classification for VHR data are mainly centered on three aspects: (1) The intra-class variability and inter-class similarity affecting classification accuracy; (2) imbalance, inconsistency and lack of quality training data for training high-accuracy classifiers; (3) large domain gaps across different scenes and geographical regions when scaling up the well-trained classifiers from a particular dataset.

1.2.1. Intra-Class Variability and Inter-Class Similarity for VHR Data

The VHR data at a resolution of a meter or less on the ground, have brought the benefits of obtaining greater details for earth observation. Along with the increased resolution, it has introduced greater intra-class variability and inter-class similarities: if using the spectral information alone, a pixel may be easily identified as belonging to multiple landcover classes; equivalently, different classes may contain pixels with similar spectral signatures [16]. Although solutions for this particular challenge such as object-based methods or spatial–spectral features [17,18] were intensively investigated, the achievable improvements failed to keep up with the increased resolution and volume of data. As a result, it may become even more problematic when advanced (and more complex) models with increased annotated datasets are used. For example, deep-learning (DL) models bring in drastically improved accuracy for specific and well-defined tasks, as they tend to fit the varying signals of the same class and at the same time to discern the tiny difference of classes with similar spectrum information (e.g., roof vs. ground). However, as is well known, this is at the expense of generalization, which may drastically decrease the performance when applying a trained DL model to different datasets with drastically different data distributions [19,20].

1.2.2. Imbalance, Inconsistency, and Lack of Quality Training Data

The increasing volume of VHR data and complexity of models naturally demand more training data, while the traditional manual labeling approaches primarily used in the era of processing coarse resolution data (such as MODIS, Landsat, and Sentinel) [21–23] or VHR but within small AOI, are sub-optimal and no longer feasible as the models are transitioned to the more data-demanding DL models. To overcome this issue, researchers sought to collect training samples from multiple sources, including crowdsourcing services (e.g., Amazon Turk) [24], and public datasets (e.g., OpenStreetMap) [25]. These additional datasets, on one hand, possess a great effort-reducing value for training high-accuracy classifiers, while on the other hand, introduce additional challenges that may need solutions for common training data issues detailed in the following.

Imbalanced training samples: Imbalanced training samples are often associated with the scene, as the number of training samples per class might not be necessarily the same

and can be scene dependent. This was somehow inherently handled in traditional manual labeling approaches, as samples were purposefully drawn and post-resampled for shallow classifiers. For DL-based models, often all available training data are fed into a network regardless of their balance; therefore, appropriate strategies both in data augmentation, training and testing are required to accommodate the imbalanced training data problems.

***Inconsistency of training samples*:** There has been a recent boost of crowdsourcing or public datasets for researchers in the RS community to perform semantic segmentation [26–28]. However, these crowdsourcing datasets or public benchmark datasets may come with inconsistent class definitions and the level of details. For example, some datasets consider buildings as part of the urban class consisting of all man-made objects on the ground, while some others consider a more detailed classification that separates buildings from other man-made structures (an example is shown in Figure 1a); some datasets define the ground class as an inclusive class that contains low-vegetations, grass, and barren land, and some others separate the ground into range-land with low-vegetation, and barrens and agricultural field as separate classes (an example is shown in Figure 1b). Therefore, the first challenge to harness the use of such data is on how to adapt or refine their labels to fit specific needs and details on the classification tasks.

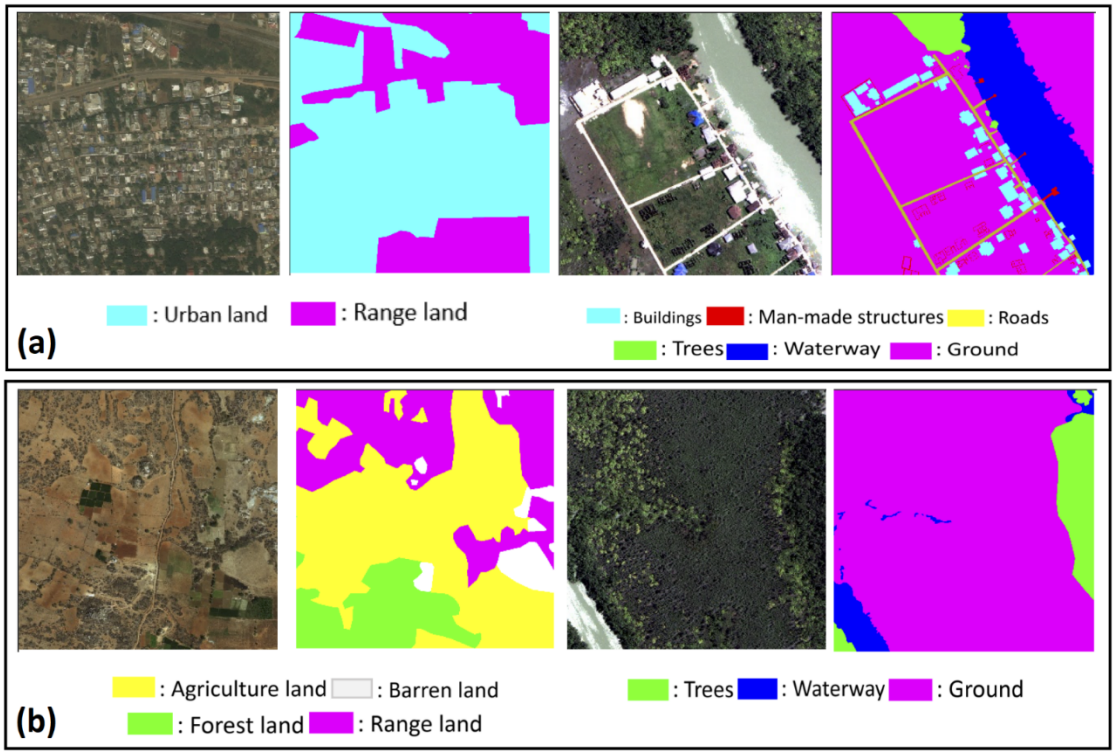


Figure 1. Sample image and label patches from publicly available benchmarks showing inconsistencies of the class definition and level of details, leading to challenges of using them directly as training sets for various RS classification tasks. In (**a**), different benchmark dataset shows different levels of details (Left: A sample patch from DeepGlobe dataset [27]; right: a sample patch from the DSTL dataset [29]) and (**b**) different benchmark dataset shows different class definitions (Left: A sample patch from DeepGlobe dataset [27] with no specific definition of the ground; right: A sample patch from DSTL dataset [29] defines the ground class but include many low vegetations and, inaccurately, some high vegetations).

Lack of quality training data: As indicated in [30], the accuracies of the most existing machine learning models in RS are underestimated, often as a result of being polluted by imperfect and low-quality training data. This presents as a common issue although

active efforts are being taken e.g., to feed the community more data as samples, thus the low-quality data assumed for learning algorithms can present another challenge.

1.2.3. Model and Scene Transferability

Transferability is often a desired feature of trained models, i.e., even the test data are captured from different sensors or different geographical locations with different land patterns compared with training data, the model will yield satisfactory performance as though it were applied to the source dataset (where the training data were collected). However, the domain gaps in RS images are often underestimated [30]. Many computer vision applications claimed good transferability at the task level [31], for example, semantic segmentation or depth estimation of outdoor crowdsourcing images [32,33] are regarded to generalize well, which are somehow inherently determined by the structure of the scene from a ground-view image: the lower part is ground, left and right sides are the façade of buildings or road extended to skylines and the upper part of the images are mostly sky. Whereas in RS images, the content of different parts of the images may come with a large variation and thus are completely unstructured, in addition to which the atmospheric effects create even larger variations on the object appearances, let alone the drastic change of land patterns across the different geographical region (e.g., urban vs. suburban, tropical area vs. frigid area, etc.). It is well-noted that every single RS image could be a domain [34,35]. Therefore, to scale up classification capabilities, transferability issues remain one of the main challenges to face.

1.3. Efforts of Harnessing Novel Machine Learning Applications and Multi-Source Data under the RS Contexts

The above mentioned challenges represent the major barriers in modern VHR RS image classification. In addition to enhancing model performances, there have been efforts that tend to utilize multi-source/multi-resolution data and unlabeled data, as well as more noise-tolerant models and learning methods to address these challenges. In general, these efforts can be collectively summarized as below:

(1) ***Weakly supervised/Semi-supervised learning to address small, imprecise, and incomplete samples.*** Weak supervision assumes the training data to be noisy, imprecise, and unbalanced, while cheap to obtain (e.g., publicly available GIS data). The approaches are often problem-dependent since it aims to build problem-specific heuristics, constraints, error distributions, etc., into the learning models. In RS, this is related to applications dealing with crowdsourcing labels, or labels that have minor temporal differences; semi-supervised learning assumes a small labeled dataset and the existence of a large amount of unlabeled data, the goal of which is to learn latent representations based on the labeled data and uses the limited training data to achieve high classification performance. Special problems or cases of semi-supervised learning include "X"-shot learning (e.g., zero-shot, one-shot, and few-shot), which needs to specify the amount of available labeled data [36,37].

(2) ***Transfer learning (TL) and domain adaptation approaches to address domain gaps.*** TL is defined in the machine learning domain that assumes knowledge learned from one task could be useful if transferred to another task. For example, a model that learns to perform per-pixel semantic segmentation of scenes can improve human detection. In the RS domain, this largely refers to techniques that minimize gaps in the feature space to achieve a generalizable classifier for data of different sensors or of different geographical locations.

(3) ***Use public GIS data or low-resolution images as sources of labeled or partially labeled data***. OpenStreetMap offers almost 80% of GIS data coverage of the globe with varying quality [38], and some local governments release relatively accurate GIS data for public distribution. Researchers had showcased work under this context and achieved conclusions specifically tied to datasets. In addition, as the low resolution labeled data with global coverage are becoming gradually more completed (e.g., National Land Cover Database [1]), these low-resolution labels can be used as a guide to generally address

domain gaps of data across different locations for scaling up the landcover classification of VHR data.

(4) ***Fusion of multi-modality and multi-view data***. Frequently, there are multiple data sources such as LiDAR (light detection and ranging), SAR (Synthetic Aperture Radar), and nighttime light data. Although these data are mostly unlabeled, they provide additional sources to explore heuristic information and more robust latent representation learning.

2. An Overview of Typical Landcover Classification Methods

The RS community has been experiencing a period where DL techniques are more frequently being used, and continue to create "new norms" regarding the workflow to produce various geospatial data products at scale. The conventional dichotomous view of landcover methods that categorizes these methods with either a pixel-based or object-based approach, designated for handcrafted features in traditional (and previously dominant) classifiers such as support vector machines and random forests, may no longer serve these "new norms" and require expansions to accommodate the DL techniques in landcover mapping. The emergence of DL techniques significantly changes the landscape of the technical field for processing high\very-high resolution images. Four primary categories of DL models have been used to extract information from RS images, including (a) scene classification that classifies an image patch and produces a scalar value representing one of the scene types, (b) semantic segmentation that takes image patches as the input and obtains the landcover types for each pixel in the image patch, (c) object detection that processes the image patch to detect object type and produces bounding boxes for each detected object, (d) instance detection and segmentation which not only provides a bounding box as performed by object detection but also delineates the boundary of the object within the bounding box as done by semantic segmentation. Readers are encouraged to refer to [8] for an overview of applications of these four categories of DL models in the RS area. In terms of landcover mapping, scene classification and semantic segmentation models are the most relevant to landcover classification methods. It should be noted that the results of the instance detection and segmentation models may also serve for landcover mapping since their predictions include the semantic segmentation results.

The DL classifiers can differ from traditional classifiers in both input and output formats. The DL classifiers combine the feature learning, feature extraction, and target classification in an end-to-end manner, and take an image patch as input rather than the one-dimensional hand-crafted feature vector required for traditional classifiers such as SVM and RF. Moreover, depending on the type of DL classifiers, the DL classifier can produce a 2D array as an output instead of a scalar value as generated by the traditional classifiers.

In this section, we aim to present a simple way for readers to systematically review the existing landcover methods and recognize their relationships considering the analysis unit. To this end, we created Figure 2 to expand the traditional dichotomous view and to present an overview of the landcover mapping methods inclusive of DL methods. In this figure, the grid icon in the first row represents the RS images in raster format and for simplicity, the band dimension is not shown. The oblique grid denotes the feature extraction area and the purple color represents the pixels to which the classification results are assigned; details are explained in the following subsections.

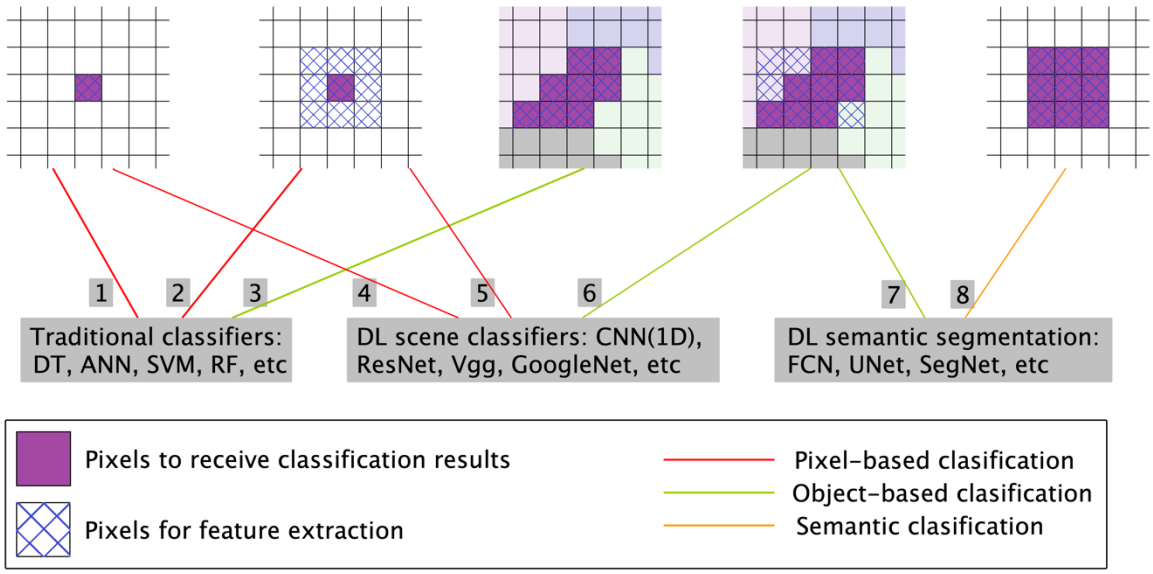


Figure 2. An overview of existing land cover classification paradigms in terms of their analysis unit. Purple color represents the classification result assignment area, oblique grid denotes the feature extraction area and different colors in the third and fourth grids represent different objects in the object-based image analysis. Different paths connecting different categories of classifiers and grids are numbered and will be referred to as Path-# when discussed in text.

2.1. Pixel-Based Mapping Method

The pixel-based method treats a single pixel as an analysis unit. The simplest implementation of the pixel-based method is shown by Path-#1 in Figure 2, which utilizes the pixel values for all the bands in this pixel location (purple color indicates pixels of concern) to form a feature vector, and employs the traditional classifiers such as random forest (RF) [39,40], support vector machine (SVM) [41,42], artificial neural network (ANN), and maximum likelihood classification (MLC) to perform the classification [43]. In this scenario, one feature vector leads to one scalar value (often an integer), which represents one landcover type as a classification result and is assigned to this pixel location. This procedure is applied to all pixels to produce the landcover map. It should be noted that the bands are not limited to original optical bands, and features such as optical index (e.g., NDVI), and ancillary data (e.g., Digital Height/Surface Model) can also be stacked to the optical bands to derive the feature vectors. To utilize the contextual information for classification, the feature extraction area can be expanded to the neighborhood area of the pixel under consideration as shown by the oblique grid in the raster associated with Path-#2 in Figure 2. In this case, additional features (e.g., standard deviation, gray level co-occurrence matrices [44], etc.) based on the pixels within the neighborhood area can be extracted to represent the texture information of the area surrounding the pixel. The neighborhood area is usually represented by a window centered on the pixel and the window size was shown to be important in impacting the classification results [45]. Readers are encouraged to refer to [46] for more information regarding the pixel-based classification.

DL scene classifiers (e.g., ResNet [47], AlexNet [48], VGG [49], GoogleNet [50], DenseNet [51], SqueezeNet [52], ShuffleNet [53], Inception [54], MobileNet [55]) use image patches as input and output a scalar value as a classification result. The window centered by a pixel can be readily used to obtain an image patch. Therefore, scene classifiers have been commonly used to perform pixel-based classification (Path-#5 in Figure 2). As convolutional neural networks (CNN) rely on the convolutional kernel to slide over the image to perform convolution operations to extract features, the relatively small window (e.g., 3×3 or 5×5 etc.) used for traditional classifiers to extract features may act as a local feature extractor that is unsuitable for CNN to extract features globally. Hence, the window used

for DL classifiers to perform pixel-based classification is comparatively larger, and various sizes of windows were used in existing works, e.g., 5 × 5 [56], 9 × 9 [57], 15 × 15 [58], 400 × 400 [59]. To reduce the computational burden and noise in the resulting classification map, a block of pixels (e.g., 3 × 3, 5 × 5) rather than a single pixel in the window center may be used to assign the classification result [59]. This is a common strategy for landcover mapping tasks that do not require strict delineation of target boundaries (e.g., human settlement mapping). A drawback of this approach is the reduced resolution of the landcover map, since all pixels in the block share the same class label.

Due to the 2D nature of the convolutional kernel in CNN, CNN usually does not apply to single pixels. However, for hyperspectral data with hundreds of bands, the standard 2D convolutional filters can be adapted with a 1D convolutional kernel to take a 1D vector as input (Path-#4 in Figure 2) [60]. In addition, the long feature vector generated from a large number of hyperspectral bands can be folded to form a 2D array, which can be input into standard 2D CNN to perform classification [61]. Readers may refer to [4] for techniques of processing hyperspectral images using DL classifiers.

2.2. Object-Based Image Analysis (OBIA)

While the pixel-based method is straightforward to implement and remains popular for medium resolution RS images [41], it introduces undesired effects for high\very-high resolution (VHR) RS images (e.g., salt-and-pepper effects). This is due to the content in a single pixel capturing incomplete information of ground objects. To overcome this issue, object-based image analysis (OBIA) was introduced. The word "object" is somewhat misleading, since it usually does not correspond to one complete object in the real world, e.g., a footprint of a complete building or a tree crown. Instead, one object used in OBIA is simply a group of homogenous pixels, representing the subarea of a landcover class or one real object, thus sometimes these are alternatively termed region-based or segment-based classification. It must be noted that an object is also referred to as a super-pixel among the computer vision community. Objects or super-pixels are generated by image segmentation algorithms. In the RS community, multi-resolution segmentation provided by the eCognition package is usually used for performing OBIA and recently open-source algorithms such as Quickshift [62], and SLIC [63] have been increasingly used. Recently, the Simple Non-Iterative Clustering (SNIC) [64] segmentation algorithm has become extremely popular when OBIA approaches are implemented in the Google Earth Engine. In Figure 2, green paths represent object-based approaches, and different colors in the grid denote different objects.

OBIA assumes that all pixels within an object belong to one single landcover class. This assumption has implications on two aspects regarding the OBIA procedure: the first one is that features are extracted from the object instead of single pixels, since object-based features are assumed to be more informative than the features extracted from single pixels; the second implication is that all pixels within the object will be assigned the same predicted label. Pixel- and object-based methods were compared and systematically analyzed in the RS community, and the consensus reveals that OBIA generates a more visually appealing map with similar or higher accuracy compared with pixel-based methods for high\very-high resolution images [65], while some studies showed that OBIA did not show advantages over pixel-based methods using medium resolution images (e.g., 10 m resolution SPOT-5) [66] in terms of either accuracy or computational efficiency.

Traditional classifiers rely on hand-crafted features extracted from pixels within objects for classification. As compared to the pixel-based method using a neighborhood area to extract features, the object-based method allows the extraction of geometric features (i.e., the features characterizing the shape of objects, such as area, perimeter, eccentricity, etc.), even though those features were less useful in improving the accuracy for supervised classification in some studies [67]. Path-#3 in Figure 2 represents the object-based approach using traditional classifiers, labels are assigned in the area where features are extracted. Readers are encouraged to refer to [68] to review the progress of OBIA.

DL scene classifiers require image patches as input, which can be naturally generated by cropping the image based on the bounding boxes of objects. Therefore, the object-based classification problem is converted to a scene classification problem. This implementation of the object-based method using CNN is shown in Figure 2 through Path-#6, whose associated grid image indicates that the feature extraction and label assignment area are not necessarily equivalent. It should be noted that since the objects are different from each other, the image patches derived from bounding boxes of those objects may have different dimensions. These image patches must be normalized to share the same dimension as specified by the input dimension of DL scene classification models.

The object-based CNN is not limited to the implementation mentioned above and several other types of object-based CNN implementation exist. For example, instead of generating one image patch for an object, users can generate multiple image patches within the object, where the size and quantity of the generated image patches are guided by the shapes and areas of the object, and majority voting was used to summarize all the classification results produced by the image patches within the object to produce the final classification result [58,69–71]. Moreover, the OBIA can be used in a post-processing manner for classification to remove salt-and-pepper artifacts. For example, CNN is firstly implemented within pixel-based approaches (i.e., Path-#5 in Figure 2) to generate a landcover map and then objects are overlaid onto the landcover map, and the label is assigned based on the majority vote, to reduce noise and yield a smoother landcover map [56,72–74]. A similar post-processing strategy via objects can be also applied to maps generated by semantic segmentation models [72,75]. OBIA became popular since its purposeful introduction to address landcover mapping tasks with VHR images [68,75–85], but we observe that the DL semantic segmentation approach, as described below, shows a tendency to replace the OBIA for VHR classification in the future.

2.3. Semantic Segmentation

The standard implementation of the semantic segmentation model takes the image patch as input and generates a label for each pixel in the image patch. The procedure corresponds to Path-#8 in Figure 2, and the associated grid image shows that the feature extraction and result assignment are on the same image grid. Semantic segmentation aims to provide a dense and per-pixel label prediction for the image patch; thus, the label assignment or feature extraction is not regarded as dependent on any single pixels or objects, but rather a dense label grid as a whole.

Landcover mapping using semantic segmentation requires a tiling approach; the large-sized RS images are split into overlapping or non-overlapping rectangular image patches. Based on this, labeled image patches are generated separately using a trained model and then stitched back to form a full classification map [86–93]. Given its superior performance in practice, the semantic segmentation models are becoming the most attempted in the RS community using VHR images to generate landcover mapping at scale. For example, building footprints, road network maps, landcover maps, and human settlement maps were successfully generated using the semantic segmentation model. The drawback of the semantic model, as compared to scene-level classification, is that it requires densely labeled training samples, which can be made very expensive. However, utilizing a patch-level label with semantic segmentation models is possible. For example, ref. [79] an assigned background as a class type to all pixels that were outside the object boundary within the image patch, with which a method to train the semantic model using the samples that were collected on object level was presented (Path-#7 in Figure 2) and showed better performance than the scene classifiers. Readers may refer to [94] for more information regarding landcover mapping using semantic segmentation models.

3. Literature Review of Landcover Classification Methods Addressing the Data Sparsity and Scalability Challenges

Both the traditional and DL methods for landcover classification require that the annotated training samples are somewhat similar (or geographically close) to the images, which is however difficult to meet, since data of varying geographical regions present vastly different land patterns that are impossible to encapsulate within one single training dataset. To address such challenges, approaches either originated from the machine learning/computer science domain, or the RS domain, were developed to overcome those challenges specific to RS image classification. Here we generally review these efforts under three related/partially overlapped areas: (1) weakly/semi-supervised classification; (2) transfer learning and domain adaptation; (3) multi-source and multi-view image-based classification. Note that we do not intentionally treat traditional and DL classifiers differently, rather we regard them (the classic and deep learning) as models with different complexity. These methods addressing data sparsity and scalability challenges, when applied to RS scenarios, can be briefly described in the following table (Table 1).

Table 1. A summary of various learning approaches addressing data sparsity and scalability challenges.

Methods	Descriptions	Application Scenario in RS Data	Examples of Relevant Works
Weakly supervised/ Semi-supervised learning	Semi-supervised learning aims to address tasks where a small set of labeled data and a large amount of unlabeled data are available, while Weak supervision assumes the labeled data to be noisy and contain errors, and the learning methods consider the uncertainty level of the available label information. In RS, this is often mixed-used with semi-automation. The readers may refer to the explanations in the texts	In RS classification, the noisy inputs are categorized as the following three types: (1) Incomplete: collected samples are too few and biased. (2) Inexact: the form of the training sample does not match with the desired form of classification results. e.g., point samples or scene-level samples vs, per-pixel samples. (3) Inaccurate samples: error-prone training samples, such as those from crowdsource data (e.g., OpenStreetMap)	[20,36,74,95–116]
Transfer learning and domain adaptation	Transfer learning (TL) is defined as transferring learned knowledge from one task to the other, normally by understanding the distribution of the feature space are different and need to be aligned through domain adaptation methods.	In RS, TL is normally defined as transferring knowledge (e.g., for classification) learned from one dataset and applied to another dataset that is drastically different in geographical location, or captured by different sensors/platforms. This also includes cases where deep models need to be fine-tuned given sparsely labeled data for training.	[20,117–126]
Multi-Modal and Multi-view learning	Data fusion methods are general approaches that utilize multiple coherent data sources or labels for performing classification tasks. Multi-view image-based learning is a subset of data fusion approaches that utilize the redundancies of multi-angular images to enhance the learning and is less covered in the literature, which this section will focus on.	Data fusion approaches are widely applicable since multi-modality remotely sensed such as SAR, optical, and LiDAR data, as well as multi-resolution, multi-sensor and multi-view data. The use of multi-view/multi-angular data is very common in photogrammetric collections. Using multi-view images enhances augments information of an area of interest and hence improves the accuracies.	[80,127–130]

3.1. Weak Supervision and Semi-Supervision for Noisy and Incomplete Training Sets

Weakly supervised methods refer to training paradigms that consider training data as noisy inputs. Such "noisy and incomplete inputs" [95] can be categorized into several cases: (1) incomplete: small training sets that are unable to cover the distribution of testing sets; (2) inexact: training sets and labels do not match with the testing sets, e.g.,

different resolution and details for labels, such as scene-level labels versus pixel-level labels; (3) inaccurate: training sets and labels are not trustworthy and contain errors. In the RS field, weak supervision under the case of "incomplete inputs" is often mixed with semi-supervision, and the only minor differences are that semi-supervision assumes a large amount of unlabeled data to draw marginal data/feature distributions while weak supervision does not require so. Given the large volume of data in RS classification tasks, such minor differences are neglectable, thus we use them interchangeably under this context. These methods primarily focus on training paradigms themselves, such as data augmentation, or formulating regularizations to traditional or DL models, to avoid overfitting. In the following, we review relevant methods in addressing "noisy and incomplete" data on either one or multiple of the above-mentioned challenges.

3.1.1. Incomplete Samples

When samples are incomplete (or not representative) to characterize the data distributions, the learned classifier might be biased even for well-defined classification problems. To differentiate this from domain gap problems, here we limit our review to tasks assuming no large domain gaps, i.e., the training samples ideally capture data points within the distribution of the testing datasets. In classic RS, the representative training samples are usually assumed to be a prerequisite, while it has evolved as an emerging challenge with the data and resolution becoming larger and higher. The line of approaches in addressing this, are to intuitively generate more points through (1) class-specific information, saliency, or expert knowledge, and (2) active learning through statistical or interactive approaches, which are described in the subsections below.

Generating new samples using saliency and expert knowledge: A common scheme is to propagate samples alongside the dataset based on neighborhood similarity [96,97], or saliency maps produced with these limited samples, followed by a re-learning scheme [98,99]. The neighborhood similarity-based approach generally assumes that the connected pixels around the sparsely labeled pixels might share the same label, thus can be added to the training sets. This is particularly useful for applications with incomplete samples, for example, road central line data can be obtained through GIS databases, while per-pixel road masks can be too time-consuming to annotate, therefore, additional labels can be added based on saliency maps learned from sparse road pixels from the central lines [103]. A few other studies explore criteria deciding whether these neighboring pixels should be incorporated in the training [96,100–102], and once identified, neighboring pixels can be taken through pixel-centric window [105] or through confidence maps of the post-classification (using the sparse training data). Additionally, RS data comes with its advantage in pattern recognition, in that the multi- or hyper- spectral information provides physical-based cues on different land classes. For example, the use of well-known indices such as NDVI (normalized difference vegetation index) [104], NDWI (normalized water index) [106], BI (building index) [131,132], and shadow index [133], which provides cues of different land classes through spectral or spatial characteristics of the RS data. For example, [19] utilized a series of RS indices as cues to introduce more samples to balance the distribution of samples to yield better classification prediction. Similarly, for wide-area mapping applications, indices are used to indicate coarse and class-specific cues for stratified classification to a finer classification level [107]. Moreover, this type of approach is used in domain-specific classification problems such as those for crop or tree species detection, where vegetation indices are implemented to introduce samples or provide expert knowledge for classification [134,135].

Generate new samples through active learning: Active learning seeks to generate data points through pre-classifying the unlabeled data and incorporating data points with confidence, or to allow users to interactively find and add the most important data points that will improve model training. A common paradigm for active learning [136], is to firstly feed the sparse data to classifiers and decide new sample points to be included, based on various criteria involving the use of posterior confidences of the classifier. For example,

the unlabeled samples can be firstly classified using available samples, and based on the classification confidence, these unlabeled samples will be ranked based on the top two most probable classes, and manual interactions are needed for ambiguous pixels, i.e., top two classes have similar confidence [137]. Authors of [138] had explored this concept for RS data through both pixel-based or cluster-based strategies, and concluded that active learning can effectively improve classification accuracy. A similar approach based on such a concept can be found in [98], with incremental improvements on feature selections and adopting an iterative scheme to further improve the accuracy for classification tasks where samples are insufficient. It should be noted that since there is a human-in-the-loop component in the process, there are possibilities that new classes can be discovered.

3.1.2. Inexact Samples

Often there exist discrepancies in the form of training samples and the desired representation in the classification. For example, it might be possible that only scene-level (or chip-level) labels are available in certain RS datasets, or in some domain-specific applications such as tree species classification, only individual points marking tree plots (e.g., from Global Positioning System(GPS) points) are available, while the semantic segmentation desires labels for each pixel [105,108], thus under these cases the training samples are inexact. The basic idea for addressing this type of classification problem follows a similar idea by transforming the point-based or scene-based training sample representation to per-pixel and dense probability or saliencies, with the assumption that the scene-level annotations contain most of the contents as described by the label, e.g., a scene with a label "urban scene " is assumed to contain most of the pixels as urban pixels such as buildings or impervious ground. The salient maps are often extracted as one of the feature maps of a deep neural network (DNN), or from classification activation maps (CAM), or posterior classification confidence maps from shallow classifiers. For example, the semantic segmentation approach in [108] used a low-ranked based linear combination of feature maps trained through scene-level data masks, to decide pixel-level content associated with each class; ref. [102] used the CAM layer of a trained network (using sample point centered image patches) to identify locations of red-attacked trees, and they demonstrated that the CAM layer can identify spatial locations at the pixel-level and as a result for deciding pixel-level labels at prediction. A similar concept was tested by [105], in which U-Net [109] was trained by using patches centered around sample points, with the last layers straightened through a global average pooling to match the single-point training data. In a prediction phase, the per-pixel prediction is performed by thresholding the CAM of the last layer (the same size as the original image). In addition to the inexact representation of training and testing data, a common cause of inexact data can be the mismatch of training labels to the desired labels. As shown in Figure 1, training and testing labels may be at a different level of detail, or the testing sample might have new classes, which can be potentially addressed by zero-shot learning (assuming no label for the new class) [36] or few-shot learning (assuming only a very limited number of sample for the new class) [110,111]. These are relatively less investigated in RS: an example of the former (zero-shot learning) [35] performed a label-level propagation-based language processing models to derive new labels of unseen classes, and an example of the latter adopted the metric-learning strategies by learning a distance function using a large image database [139,140], and by implementing the distance function, it could thus perform predictions. Both of these examples have demonstrated certain levels of improvements [110,141]. Few-shot learning has been recently investigated in RS community for scene classification and showed promising results [142,143], but its application for semantic segmentation model for RS images remains relatively underexplored.

3.1.3. Inaccurate Samples

In an RS context, inaccurate samples mainly refer to cases where the samples contain a certain level of errors or inconsistencies, such as those aggregated from open or crowdsourcing data (e.g., OpenStreetMap [25]), or data labeled from low-resolution data [95].

OpenStreetMap (OSM) is one of the most investigated data sources for classification studies, as it was reported to cover approximately 83% of road networks (as of 2016) [144] and 40% of a street network [38]. Most of the existing studies assume OSM to be error-prone, thus various strategies were developed to refine the labels. For example, [74] explored a data fusion and human-in-the-loop approach, which fused extracted vector data from OSM and overlays with satellite images for visual checking and sample selection, to perform quality control of the labels before training; ref. [112] automated this procedure by using several common RS indices including NDVI [104], NDWI [106] and MBI [131], to remove inconsistent labels using heuristic rules. Authors of [113] employed a simple procedure to refine per class labels: by assuming registration errors of the OSM and the data, they first eroded the OSM labels through binary morphological operations, and in a second step, they performed a clustering procedure on pixels, and labeled them based on the OSM. Both of these studies [112,113] used shallow classifiers. It should be noted that in their studies, the road vectors were considered to be sufficiently accurate and were directly imposed on the final results. DL-based approaches applied to the OSM data [114,115], often consider that the volume of OSM for training is sufficiently large to avoid overfitting problems in DL, thus there is only very light, or no pre-processing applied to the OSM data before training. Alternatives, in addressing these data inaccuracies, the DL methods build loss functions that inherently consider data inaccuracy, such as data imbalances and label noisiness. For example, the work of [114] used pre-defined empirical weights on the losses of different classes (i.e., smaller weight for background and bigger weights for building and road classes) to address the label sparsity. The work of [115] used a similar strategy in their loss function for its application of map generation using generative models. Both works [114,115] reported satisfactory classification results in their applications. There are also other works that develop ad hoc solutions specific to the type of crowdsourcing data or the OSM data. For example, the work of [144] directly utilized OSM data as tile-level input to train a random forest classifier for predicting region-level socioeconomic factors, which reported satisfactory results without any pre-processing of the OSM; the work of [116] reported an application of using crowdsourcing geo-tagged smartphone data for crop type classification, and they reported that the data were manually cleaned for DL.

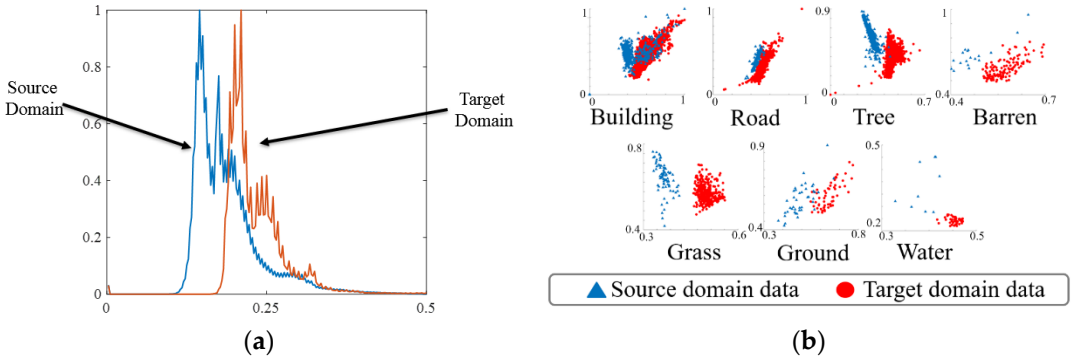
3.2. Transfer Learning and Domain Adaptation for RS Classification

As mentioned above, transfer learning (TL) in machine learning is broadly defined as transferring learned knowledge from one task to other related tasks, for example, applying information learned from the semantic segmentation task to the human recognition task [145]. In RS classification, this often explicitly refers to training a model (or classifier) using one dataset and applying the adapted model to another dataset (with no or very few labels). This is particularly needed in the context that the two datasets stem from different geographical regions, or are collected from different sensors and platforms [18]. In this case, feature distributions extracted from the dataset are different, and it requires domain adaptation methods to minimize the differences of the features across different datasets. In RS classification, the domain with labeled data is defined as the source domain, and the domain with no or very few labels (as compared to the source domain) is regarded as the target domain. There are two types of TL applications when applied to RS applications: (1) domain adaptation and (2) model fine tuning.

3.2.1. Domain Adaptation

Domain gaps are defined as the differences of feature distributions between source and target datasets. This is often regarded as the major cause of the generalization issues among machine learning algorithms. Figure 3 illustrates a simple example that draws the distribution of radiometric values of different datasets: Figure 3a draws typical marginal distributions of the radiometric values of two datasets, which shows a systematic bias; Figure 3b draws the conditional distributions (per class feature distribution) of a two-dimensional feature, which shows that, for the same class, the feature values appear

differently. These differences made it different for a classifier trained from the source data, to achieve satisfactory results in the target data. The domain adaptation (DA) methods aim to minimize such differences between feature distributions, which can be either performed explicitly on the feature level (through both handcrafted or learned features), or implicitly performed in latent spaces during the learning process. In the following subsections, we introduce a few typical DA methods used in RS. (1) Feature-level domain adaptation; (2) Domain adaptation through DL-based latent space; (3) Generative adversarial Network (GAN) [146] based domain adaptation.



(**a**) (**b**)

Figure 3. An intuitive visualization of typical domain gaps and misalignment of feature distributions between the source and target domain: (**a**) Marginal distribution (feature distribution over all the classes) of radiometric values; (**b**) conditional distribution of a typical two-dimensional feature of the two domains showing the misalignment on different object classes, x and y axis of these subfigures plots are appropriately adjusted for best visualization.

Feature-level domain adaptation: The feature-level DA refers to approaches that minimize the feature distributions between the source and target domain. DA approaches may minimize either marginal or condition feature distributions, or both, depending on the availability of these distributions in the target data. Normally aligning the marginal distributions assumes the algorithm is completely agnostic to label space of the target domain, while aligning the conditional distributions requires part of the labels of the target domain data to be known. A DA algorithm proposed by [147] was regarded as one of the simplest methods, which produced copies of features based on limited labeled samples from the target domain, to improve the accuracy for target domain classification, which reported to have achieved competing results in many classification tasks. Since this simple algorithm requires some labeled data in the target domain, ref. [117] proposed another "simple" DA algorithm termed CORAL (CORrelation ALignment) that does not require any labeled target domain data, as it aligns the marginal distributions of the features in the source and target domains through minimizing the differences of the second-order statistics of the distributions.

An important line of work in DA, known as reproducing kernel Hilbert space (RKHS) [118] aims to project both the source and the target feature space to a common metric space. In RKHS, the differences of the feature distributions are minimized under a metric defined as maximum mean distance (MMD), and following the minimized distributions, the classification of the target data can be then performed in the new metric space. An improved version, called transfer component analysis (TCA) [119], incorporated feature transformation (so called transfer component) into the MMD minimization, thus yielding better results. Its semi-supervised version was proposed in [119], which minimizes the conditional distribution by assuming only very limited labels in the target domain. However, it was known that minimizing either marginal or the conditional distribution alone does not encapsulate the entire data distribution. Thus, ref. [120] proposed a joint domain adaptation (JDA) method that jointly optimizes the alignment of the marginal and conditional distribution of features, in which the features were transformed under a

principal dimensionality reduction scheme. While requiring a longer computational time, JDA appeared to perform better than TCA.

When target domain labels are not available, an intuitive scheme for generating so-called pseudo labels can be applied: It firstly trains a classifier on the source domain data, then performs classification on the target domain data, and finally takes labels with high confidences as the desired pseudo labels from target domains. However, this will inevitably introduce the sample imbalance problem. Researchers [19] proposed a multi-kernel approach that tends to provide the flexibility to weigh multiple kernels to improve the DA in their experiments, they specifically noted this data proposed to balance the target domain labels for training. Different from most of the aforementioned methods which are mainly practiced in tasks other than RS image classification, the work of [19] was applied to the RS contexts and reported an improvement of 47% on combined multi-spectral and DSM (digital surface model) data. Researchers [121] evaluated the semi-supervised TCA (SSTCA) method on a set of RS images, and reported that the SSTCA method, when combined with a histogram matching between source and target images, achieved approximately 0.15 of improvement in terms of Kappa metric.

Domain adaptation through DL-based latent space: Most of the aforementioned or earlier methods are designed for traditional machine learning pipelines. There were a few new developments in recent years focusing on domain adaptation for DL at the latent space or embedding vectors. Since DL methods for classification / semantic segmentation require densely labeled images, it will become more challenging to collect even a small number of samples in the target domain, thus the DA method for DL usually assumes no target labels. The basic idea of this type of approach is to obtain the source domain image, target domain image, and the source domain labels as input for training, and build a loss to infer the embedding vectors. These embedding vectors are often shared between the source and target domain data, to implicitly learn domain adaptations in the embedding space. An example of such a method is the DL version of the CORAL algorithm [122], in which the CORAL loss was built to minimize the second-order statistics of the feature maps, resulting in a shared embedding space for feature extraction and classification. Researchers [123] built the loss function as discriminators of the source and domain data in an embedded space consisting of both learned CNN (convolutional neural networks) and wavelet features, such that the embedding space is less semantic to source and target domain data. Research by [148] attributed the domain gaps to be primarily the scale differences; they built a scale discriminator to serve as a trainable loss to be scale-invariant to objects, which infer the gradients back to the embedding space to improve the DA. A similar but slightly more agnostic approach obtained synthetic data for semantic segmentation in an automated driving scenario, in which they trained a discriminator to identify whether the output semantic maps came from a synthetic or real image. This discriminator essentially serves as a loss to infer the parameters in the shared embedding space, therefore, to improve the adaptation of the semantic segmentation tasks.

GAN-based domain adaptation: The GAN-based DA methods aim to use generative models that directly translate target domain images to the styles in the source domains. The ColorMapGAN [124] presented such an approach, which assumes no labels from the target domain and turns target domain images into the images of the original domain through a CNN-based GAN, and showed significant improvement in classification accuracy. Since this type of approach is agnostic to classifiers, it works for both traditional and DL models. The authors expanded this work to a more general framework called DAugNet [33], which incorporated multiple sources and target domains, and can consistently learn from examples. Researchers [125] extended the image translation concepts to include four images as the input for semantic segmentation: the target image, target-styled source image, source image, and reconstructed source image from the target-styled source image, in which the latter three associated with the source had labels to backpropagate gradients for training a shared encoder-decoder. In summary, since this type of approach requires a heavy training

process for each pair of source and domain, it may face limitations for real-time applications during test time.

3.2.2. Model Fine-Tuning

Fine-tuning is a specific class of approach popularized along with DL methods, which tends to retrain part of a pre-trained network (e.g., ImageNet [149]) using relatively fewer samples. It has become standard practice to (1) adapt classification tasks with few samples; (2) utilize well-pretrained networks as feature extractors. Authors of [31] performed a deep analysis on the transferability of deep neural networks (representatively forward feed networks in this work), and concluded that initializing the networks with pretrained parameters and leaving any number of layers to be trained (depending on the available samples), turned out to be beneficial to achieve high classification accuracy. As a standard practice, this was applied in many RS classification tasks given the lack of labeled data. For example, in the ISPRS semantic segmentation challenge for RS images [150], the majority of the DL approaches use pre-trained networks (often from ImageNet) as a start, and achieved an accuracy beyond traditional approaches. Researchers [126] fine-tuned networks for satellite image classification, and concluded that fining-tuning half of the layers might achieve the same level of accuracy as fine-tuning the entire network, while it could obtain a higher convergency speed. This might be due to well-trained parameters that might serve as a regularization when training models. Other successful applications include (but are not limited to) the Function Map of the World challenge [151], in which solutions for scene-level RS classifications achieved the best results using fine-tuned networks. Authors of [152] used an ensemble of CNN, the training of which started with pre-trained weights and achieved 88.5% in the ISPRS benchmark. Despite the existing work having largely explored fine-tuning as an option to adapt pre-trained networks to address learning tasks with sparse labels, the detailed fine-tuning strategies, for example, freezing how many and which layers, and what regularization constraints to use, are still following a trial-and-error approach. Yet a thorough, and systematic study on fine-tuning approaches for RS applications remain lacking.

Moreover, fine-tuning with labeled samples can follow a self-supervised representation \metric learning procedure, where positive and negative pairs are generated automatically from a large amount of unlabeled images and are used to train a deep learning backbone (e.g., ResNet) to minimize the feature distance of positive pair samples and maximize the feature distance for negative samples [153,154]. Researchers [155] proposed a metric learning model named discriminative CNN(D-CNN) to perform metric learning and obtained state-of-the-art performance on three public RS datasets for scene classification tasks, showing promising results of using self-supervised learning approaches for landcover mapping task. However, it remains to be seen whether self-supervised could also help semantic segmentation using RS images, although this has been confirmed to be beneficial in computer vision community [156].

3.3. Multi-Sensor, Multi-Temporal and Multi-View Fusion

The earth observation (EO) data have been growing rapidly owing to the increased number and types of earth observation sensors launched in the past years. The large volume of EO data is heterogeneous in terms of temporal resolution, spatial, spectral resolution, collection angle, or sensor types. In addition to EO sensors, geospatial data is being generated from other sources such as social media and mobile phone locations. Investigating methods to fuse the geospatial data collected across diverse sources or different temporal steps, has attracted increasing attention from the RS community, owing to their complementary characteristics or synergistic effects which can potentially improve the accuracy, temporal resolution, and spatial resolution for the land cover classification results. While a strict definition of taxonomy regarding the data fusion approaches cannot be found in existing literature, data fusion for classification is generally considered at the pixel level, feature level, or decision level, as shown in Figure 4. In Figure 4, two sources

are used for simplicity to represent multi-modal geospatial data (e.g., sentinel 1 vs. sentinel 2, spectral vs. LiDAR, EO data vs. social media data, collection angle 1 vs. collection angle 2, collection time step 1 vs. 2, etc.), and it should be noted that the fusion approaches may be extended to multiple sources.

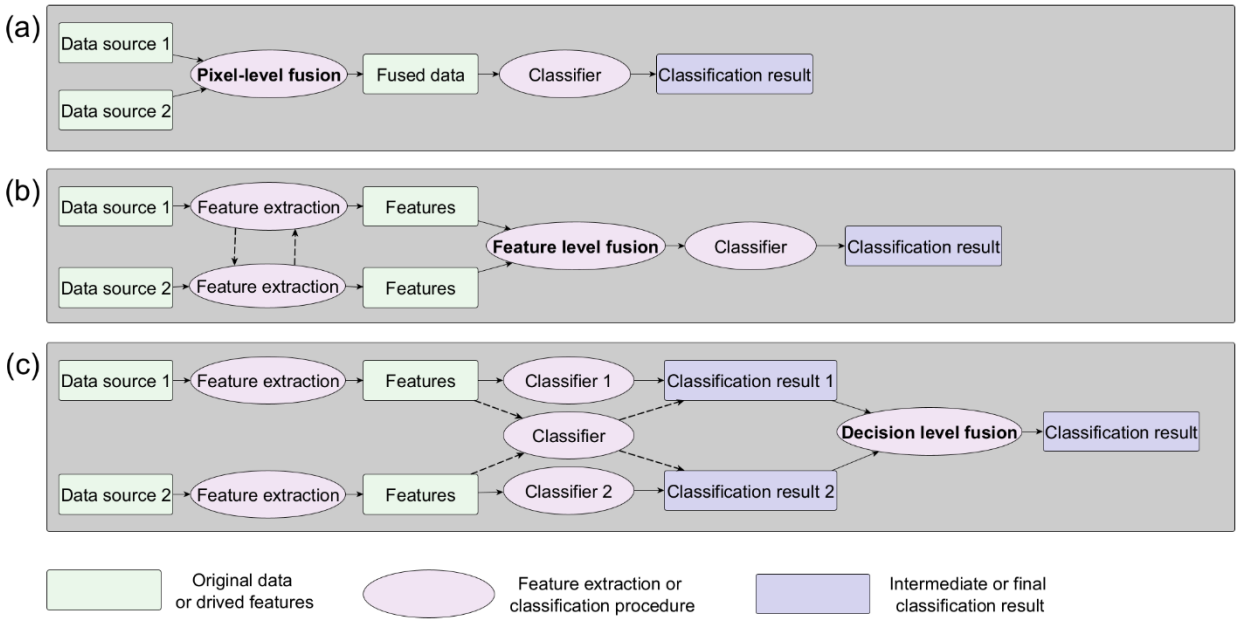


Figure 4. Data fusion for classification can be performed at the (**a**) pixel level, (**b**) feature level, and (**c**) decision level. Dashed lines represent alternative paths in contrast to solid lines.

3.3.1. Pixel-Level Fusion

Pixel-level fusion refers to the fusion of information from two sources to generate a new set of data and each band of the new set contains the information from both sources. One pixel-level fusion approach is the well-known principal component analysis (PCA): [157] performed the principal component analysis (PCA) for the Sentinel optical images, replaced with last two PCA components with Sentinel SAR images, transformed the new set of PCA components back to the original format of the optical images as the result of fused data, which was then input into the UNet [109] for crop-type classification. Another category of pixel-level data fusion method relies on pan-sharpening techniques specifically designed to fuse the high spatial resolution, single-band panchromatic with low-spatial resolution multi-band images to generate high-spatial resolution multi-band images. Authors of [158] conducted pan-sharpening algorithms to generate high resolution multiple spectral images, which were used for ice-wedge polygon detection with Mask R-CNN [159]. Eight pan-sharpening algorithms were compared in this study and it was concluded that the performance of the pan-sharpening algorithm was scene dependent, and indicated that the fusion algorithms that preserve spatial characteristics of the original PAN imagery benefit the DL model performances. Researchers [160] proposed a hybrid fusion approach that not only performs pan-sharpening to generate high spatial multi-band images, but also proposed a module named correlation-based attention feature fusion (CAFF) to fuse the features from both sources for image classification. Pan-sharpening itself is an active research topic, for which there are a few works [161–163] providing comprehensive reviews. In addition to PCA and pan-sharpening techniques, other approaches were developed. For example, ref. [164] proposed a new approach that constructed K filters corresponding K data sources to convert the K sources of data to a fused dataset with K bands called latent space, which was then used to map local climate zones with SVM classifier.

In addition to pixel-level fusion that operates in the spatial–spectral domain, it can also be conducted in the spatial–temporal domain, which primarily aims to increase the

temporal resolution of high spatial resolution images [35,165]. For example, ref. [165] developed an approach based on deep learning techniques to generate temporally dense Sentinel-2 data at 10-m resolution with information of Landsat-2 images. Temporal–spectral fusion can be decoupled with the landcover classification task and is beyond the scope of our study for a detailed review. Readers can refer to a review article on this topic [166] for more information.

3.3.2. Feature-Level Fusion

Feature-level fusion is the most common approach to fuse multi-modal data for landcover classification due to its simplicity in terms of implementation and concept. Among other types of feature fusion approaches, concatenation of handcrafted features is the most common in the existing literature. For example, features derived from LiDAR and hyperspectral images were concatenated to map land covers with random forest classifier, as is shown in the work of [167], where such approaches were applied to the Houston, Trento, and Missouri University, and University of Florida (MUUFL) Gulport datasets. Other types of features such as spectral indices, texture metrics, and seasonal metrics, can be extracted and concatenated with location-based features to map the human settlement using a random forest classifier [168].

In recent years, the number of papers that utilize the DL network to fuse the features has surged. Given datasets of two sources, a common DL architecture for fusing the features extracted from the two sources has two streams, with each stream taking one type of data as input and the features extracted from two streams in the last layer, followed by concatenation and being fed into a fully connected layer for classification. Ref. [169] adopted this DL architecture to fuse LiDAR and spectral images for landcover classification in an urban environment. To increase the feature representation, more than one type of DL network can be used to extract features from one data source. For example, long short-term memory (LSTM) and CNN were both used to extract features from social sensing signature data, which were concatenated with features extracted with CNN from spectral images to classify urban region functions [170]. The feature fusion may also be performed in intermediate layers in addition to the last layer [171–174], which is indicated by the dashed arrow lines in Figure 4b. In addition to concatenation, features can also be fused by maximum extraction operation, i.e., for each position in the feature vector, selecting the maximum values among the features extracted across all the data sources [174].

Natural Language Processing (NLP) algorithms aim to extract meaningful information from a sequence of words. One sentence consists of multiple words, and those words can be represented with a 2D array, where the length of the rows is the same as the number of words of the sentence and each row corresponds to the embedding associated with one word. Moreover, the multiple-temporal multi-spectral RS images can be represented with a 2D array as represented for one sentence with each row corresponding to the spectra values at a one-time step. Due to this connection of RS time series with NLP problems, many NLP algorithms have been adapted as a feature fusion approach to process the RS time-series data to generate the crop map. For example, the commonly used NLP algorithms Recurrent Neural Network(RNN) and Long Short-Term Memory (LSTM) [175,176] have been employed to process the 12 monthly composites of 9 bands of sentinel-2 time-series data for crop mapping task [177]. The latest state-of-the-art NLP algorithm named transformer [178] was also employed to process the time-series RS data for crop mapping [177,179]. For example, ref. [180] proposed a model that combines LSTM with a transformer attention module to process 23-time steps of a 7-day composite of 6 spectral bands of Landsat Analysis Ready Data (ARD) to generate crop data layer (CDL). The CNN module has a good capability of extracting features from the scene context, but none of those multi-temporal approaches have attached a CNN module to the sequence processing module for crop mapping.

3.3.3. Decision-Level Fusion

Decision-level fusion is performed after each classifier leads to a decision from one data source and the decision can be represented as a class label or class probability (Figure 4c). The majority voting [181,182] or the summation of class probability across different data sources [174] can be conducted to implement the decision fusion. Additionally, a final decision can be obtained based on the confidence level of the predicted class label [183]. Figure 4c shows two approaches for training the classifiers: one (solid line arrows) is to train separate classifiers for different data sources and the other (dashed line arrows) is to train one classifier using all the available sources. It should be noted the decision-level fusion used in [174] is different from the one indicated in Figure 4c since the decision vector used in [174] is more similar to the features used in feature fusion by DL networks.

3.3.4. Multi-View Fusion

The multi-view based classification method we introduce here refers to photogrammetric images or multi-view/multi-date images covering the same region. By building the relationship between the multi-view images and their respectively generated digital surface models (DSM), users can build pixel-level correspondences to share labels among these images, and at the same time utilize the spectral redundancies to improve classification results [81]. Multi-view data were considered detrimental to the classification accuracy, as multi-view images needs to be normalized to the nadir view to improve the classification accuracy [184]. However, recent studies show multi-view data contain extra information compared with single-view RS images and the redundant information contained in multi-view images can be used to improve the classification.

Early efforts employ multi-view reflectance to determine the parameters of the Bidirectional Reflectance Distribution Function (BRDF), which characterizes multi-view reflectance patterns. Then, the BRDF parameters are used as a proxy of multi-view information and are concatenated with other features or used alone as the feature vector for traditional landcover classifiers [81,185–190]. In addition to BRDF, other multi-view feature extraction methods were developed for the traditional classifiers to improve landcover mapping accuracy. For example, ref. [191] extracted angular difference among multi-view images for urban scene classification, ref. [192] applied bag-of-visual-words(BOVW) to multi-view images for urban functional zone classification, and [193] proposed Ratio Multi-angular Built-up Index (RMABI) and Normalized Difference Multi-angular Built-up Index (NDMABI) for built-up area extraction.

BRDF-based methods is computationally expensive and other methods mentioned above are tailored for specific applications with specific multi-view datasets. Researchers [182] proposed a simpler yet more efficient and general approach, which fuses multi-view objects by training the classifier using multi-view objects and obtains the final classification result by the majority-voting of inference results from multi-view objects, which works similarly as indicated in Figure 4c. This method accommodates the varied number of views and shows substantially higher efficiency for fusing multi-view information compared to the BRDF model. In addition to traditional classifiers, the method is applicable to DL scene classifiers. It was demonstrated that the convolutional neural network benefits more from the redundancies of multi-view data than traditional classifiers (e.g., SVM, RF) for improving the accuracy of landcover mapping with this method [80]. Finally, multi-view information can also be used with the semantic segmentation model. For example, ref. [114] adapted the semantic segmentation model to allow it to obtain the stacked multi-view tensors in the model for semantic mapping using multi-view information and demonstrated that their method gave better performance than methods that used different views separately.

4. Final Remarks and Future Needs

The efforts in different classes of learning methods for image classification generally originate from the machine learning/computer science domain and present a varying

degree of preferences in the RS domain in recent literature, while existing attempts in addressing different aspects of the RS landcover classification problems for very high resolution (VHR) remain limited and novel applications, adaptation, and reformulation of these contexts into RS problems are still greatly needed. We consider the landcover classification as a highly disparate problem, and on one hand, the solutions and models may need to accommodate available data and scene content, and on the other, new data sources and novel use of existing and open data may be explored for incorporation to improve landcover classification of VHR images. Considering the multi-complex nature of this problem, this paper, in contrast to other existing feature/classifier specific reviews, has provided a comprehensive review on recent advances in landcover classification practices, specifically focusing on approaches and applications that directly or indirectly address the scalability and generalization challenges: we first presented a general introduction to commonly used classifiers with an expanded definition of analysis unit to incorporate deep learning paradigms, and then described existing solutions and categorical methods in addressing the need for classification in weak/semi supervision context where samples can be incomplete, inexact and inaccurate, in addition to domain adaptation approaches in the case of model transfer and model reuse in different contexts; finally, we surveyed existing paradigms that explored the use of other types of data for fusion.

Many of the existing works using DL semantic segmentation for landcover classification, have overwhelmingly demonstrated the level of improvement. DL-based semantic segmentation methods have received increasing attention for landcover mapping, and it is likely to replace the OBIA in future, which is the current solution aiming to overcome issues brought about by VHR images but has often been criticized for imperfect results and extra computational costs of the image segmentation procedure. It is expected to continue in the future and ultimately has the potential to become the standard landcover mapping approaches. There exist benchmark datasets for methodology comparison, however, using these methods in practice are far more complicated than standard tests, due to the massive influences of the availability of quality training data, as well as the multi-complex nature of data resulting from multi-sensors and multi-modal outputs.

During the survey, we additionally discovered the consensus that using data fusion for RS classification provides a promising direction and is highly needed; with the number of relevant works continuously increasing, however, since the data sources show disparity across different works, a taxonomy and a comprehensive comparison among those methods remains currently lacking in the literature. Moreover, we noticed that researchers in RS communities do not release codes as often as those in the computer vision\science community, making the benchmarking for data fusion development in RS community more challenging. In addition, current applications of domain adaption, self-supervised training and meta-learning in the RS community have not caught up with the latest technical progress observed in the computer science community, where it is claimed that "the gap between unsupervised and supervised representation learning has been largely closed in many vision tasks" [194]. Considering unlabeled RS images are being continuously collected each day from various sensors covering different temporal and spatial domains, investigating how those unlabeled images can be utilized with self-supervised or domain adaption techniques to train a better supervised model could provide a promising direction.

It is generally agreed that solutions for landcover classification problems are still ad hoc in practice, requiring training data with good quality. However, we observe a trend where more research works are shifting their gears from achieving high accuracy with more complex DL models, to achieving more general and global-level classification. This is fueled by the need to, (1) address the challenge of data sparsity, inaccuracy and incompleteness; (2) harness the ever-growing number of sensors with different modality to achieve solutions free of moderation by experts.

To this end, based on this review, we provide a few recommendations for the future works in this research line: (1) Developing domain adaptation approaches taking advantage of the unique characteristics of RS data, such as their diversity in land patterns of different

geographical regions, the availability of low-resolution labels for semi-supervised DA, available height information globally, as well as physics-based spectrum signatures in the RS world. (2) Exploring the underlying mechanisms of spectrum diversity across different sensors, to achieve inter-sensor calibration prior to classification. (3) Alleviating the cost of training sample collection for semantic segmentation using non-standard, crowd-sourced means and developing methodologies that standardize use of common crowd-sourcing data (such as OSM) for classification, and easier means to access globally available labels for training. (4) Evaluating the extra benefits brought about towards transfer learning by big databases of labeled RS images compared with computer vision datasets. (5) Establishing more comprehensive benchmark datasets for assessing the generalization capabilities (e.g., few-shot learning task, domain adaption task) of DL solutions especially for the semantic segmentation models; (6) Analyzing the roles of self-supervised learning, active learning and meta-learning in reducing the cost of using deep learning semantic segmentation models for landcover mapping.

Author Contributions: R.Q. and T.L. both made significant contributions to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: No funding was used to support the work for this project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Homer, C.H.; Fry, J.A.; Barnes, C.A. The national land cover database. *US Geol. Surv. Fact Sheet* **2012**, *3020*, 1–4.
2. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
3. Neupane, B.; Horanont, T.; Aryal, J. Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. *Remote Sens.* **2021**, *13*, 808. [CrossRef]
4. Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *Isprs J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [CrossRef]
5. Vali, A.; Comai, S.; Matteucci, M. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sens.* **2020**, *12*, 2495. [CrossRef]
6. Griffiths, D.; Boehm, J. A Review on Deep Learning Techniques for 3D Sensed Data Classification. *Remote Sens.* **2019**, *11*, 1499. [CrossRef]
7. Bello, S.A.; Yu, S.S.; Wang, C.; Adam, J.M.; Li, J. Review: Deep Learning on 3D Point Clouds. *Remote Sens.* **2020**, *12*, 1729. [CrossRef]
8. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *Isprs J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
9. Ma, L.; Liu, Y.; Zhang, X.L.; Ye, Y.X.; Yin, G.F.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *Isprs J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
10. Pashaei, M.; Kamangir, H.; Starek, M.J.; Tissot, P. Review and Evaluation of Deep Learning Architectures for Efficient Land Cover Mapping with UAS Hyper-Spatial Imagery: A Case Study Over a Wetland. *Remote Sens.* **2020**, *12*, 959. [CrossRef]
11. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. [CrossRef]
12. Hoeser, T.; Bachofer, F.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part II: Applications. *Remote Sens.* **2020**, *12*, 3053. [CrossRef]
13. Ghanbari, H.; Mahdianpari, M.; Homayouni, S.; Mohammadimanesh, F. A Meta-Analysis of Convolutional Neural Networks for Remote Sensing Applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3602–3613. [CrossRef]
14. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
15. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
16. Qin, R. A mean shift vector-based shape feature for classification of high spatial resolution remotely sensed imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1974–1985. [CrossRef]

17. Ghamisi, P.; Dalla Mura, M.; Benediktsson, J.A. A survey on spectral–spatial classification techniques based on attribute profiles. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2335–2353. [CrossRef]
18. Fauvel, M.; Benediktsson, J.A.; Chanussot, J.; Sveinsson, J.R. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3804–3814. [CrossRef]
19. Tuia, D.; Persello, C.; Bruzzone, L. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
20. Liu, W.; Qin, R. A MultiKernel Domain Adaptation Method for Unsupervised Transfer Learning on Cross-Source and Cross-Region Remote Sensing Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4279–4289. [CrossRef]
21. Cai, S.; Liu, D.; Sulla-Menashe, D.; Friedl, M.A. Enhancing MODIS land cover product with a spatial–temporal modeling algorithm. *Remote Sens. Environ.* **2014**, *147*, 243–255. [CrossRef]
22. Williams, D.L.; Goward, S.; Arvidson, T. Landsat. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 1171–1178. [CrossRef]
23. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
24. Daly, T.M.; Nataraajan, R. Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk. *J. Bus. Res.* **2015**, *68*, 2603–2609. [CrossRef]
25. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
26. SpaceNet. SpaceNet on Amazon Web Services (AWS). Available online: https://spacenet.ai/datasets/ (accessed on 1 December 2021).
27. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
28. Schmitt, M.; Ahmadi, S.A.; Hänsch, R. There is no data like more data–current status of machine learning datasets in remote sensing. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 11–16 July 2021.
29. Kaggle. Dstl Satellite Imagery Feature Detection. Available online: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection (accessed on 24 May 2021).
30. Burke, M.; Driscoll, A.; Lobell, D.B.; Ermon, S. Using satellite imagery to understand and promote sustainable development. *Science* **2021**, *371*, eabe8628. [CrossRef] [PubMed]
31. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3320–3328.
32. Li, Z.; Snavely, N. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2041–2050.
33. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
34. Tasar, O.; Giros, A.; Tarabalka, Y.; Alliez, P.; Clerc, S. DAugNet: Unsupervised, Multisource, Multitarget, and Life-Long Domain Adaptation for Semantic Segmentation of Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1067–1081. [CrossRef]
35. Elshamli, A.; Taylor, G.W.; Areibi, S. Multisource domain adaptation for remote sensing using deep neural networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3328–3340. [CrossRef]
36. Li, A.; Lu, Z.; Wang, L.; Xiang, T.; Wen, J.-R. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4157–4167. [CrossRef]
37. Larochelle, H. Few-Shot Learning. In *Computer Vision: A Reference Guide*; Springer International Publishing: Cham, Switzerland, 2020; pp. 1–4.
38. Barrington-Leigh, C.; Millard-Ball, A. The world's user-generated road map is more than 80% complete. *PLoS ONE* **2017**, *12*, e0180698. [CrossRef]
39. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
40. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
41. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
42. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
43. Sunde, M.; Diamond, D.; Elliott, L.; Hanberry, P.; True, D. Mapping high-resolution percentage canopy cover using a multi-sensor approach. *Remote Sens. Environ.* **2020**, *242*, 111748. [CrossRef]
44. Mohanaiah, P.; Sathyanarayana, P.; GuruKumar, L. Image texture feature extraction using GLCM approach. *Int. J. Sci. Res. Publ.* **2013**, *3*, 1–5.
45. Chen, D.; Stow, D.; Gong, P. Examining the effect of spatial resolution and texture window size on classification accuracy: An urban environment case. *Int. J. Remote Sens.* **2004**, *25*, 2177–2192. [CrossRef]

46. Khatami, R.; Mountrakis, G.; Stehman, S.V. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sens. Environ.* **2016**, *177*, 89–100. [CrossRef]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
48. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
49. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
50. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 1–9.
51. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
52. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
53. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
54. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
55. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
56. Liu, S.; Qi, Z.; Li, X.; Yeh, A. Integration of Convolutional Neural Networks and Object-Based Post-Classification Refinement for Land Use and Land Cover Mapping with Optical and SAR Data. *Remote Sens.* **2019**, *11*, 690. [CrossRef]
57. Zhong, Y.; Hu, X.; Luo, C.; Wang, X.; Zhao, J.; Zhang, L. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H-2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. *Remote Sens. Environ.* **2020**, *250*, 112012. [CrossRef]
58. Martins, V.S.; Kaleita, A.L.; Gelder, B.K.; da Silveira, H.L.; Abe, C.A. Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution. *ISPRS J. Photogramm. Remote Sens.* **2020**, *168*, 56–73. [CrossRef]
59. Arndt, J.; Lunga, D. Large-Scale Classification of Urban Structural Units From Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2634–2648. [CrossRef]
60. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
61. Jia, P.; Zhang, M.; Yu, W.; Shen, F.; Shen, Y. Convolutional neural network based classification for hyperspectral data. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5075–5078.
62. Vedaldi, A.; Soatto, S. Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 705–718.
63. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
64. Achanta, R.; Susstrunk, S. Superpixels and polygons using simple non-iterative clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4651–4660.
65. Whiteside, T.G.; Boggs, G.S.; Maier, S.W. Comparing object-based and pixel-based classifications for mapping savannas. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 884–893. [CrossRef]
66. Duro, D.C.; Franklin, S.E.; Dubé, M.G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens. Environ.* **2012**, *118*, 259–272. [CrossRef]
67. Ke, Y.; Quackenbush, L.J.; Im, J. Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification. *Remote Sens. Environ.* **2010**, *114*, 1141–1154. [CrossRef]
68. Ma, L.; Li, M.; Ma, X.; Cheng, L.; Du, P.; Liu, Y. A review of supervised object-based land-cover image classification. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 277–293. [CrossRef]
69. Huang, B.; Zhao, B.; Song, Y. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [CrossRef]
70. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atitinson, P. An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sens. Environ.* **2018**, *216*, 57–70. [CrossRef]

71. Lv, X.; Ming, D.; Lu, T.; Zhou, K.; Wang, M.; Bao, H. A New Method for Region-Based Majority Voting CNNs for Very High Resolution Image Classification. *Remote Sens.* **2018**, *10*, 1946. [CrossRef]
72. Sun, Y.; Zhang, X.; Xin, Q.; Huang, J. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 3–14. [CrossRef]
73. Tong, X.; Xia, G.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [CrossRef]
74. Zhao, W.; Bo, Y.; Chen, J.; Tiede, D.; Blaschke, T.; Emery, W.J. Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and OpenStreetMap (OSM). *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 237–250. [CrossRef]
75. Mboga, N.; Georganos, S.; Grippa, T.; Lennert, M.; Vanhuysse, S.; Wolff, E. Fully Convolutional Networks and Geographic Object-Based Image Analysis for the Classification of VHR Imagery. *Remote Sens.* **2019**, *11*, 597. [CrossRef]
76. De Luca, G.; Silva, J.M.N.; Cerasoli, S.; Araújo, J.; Campos, J.; Di Fazio, S.; Modica, G. Object-based land cover classification of cork oak woodlands using UAV imagery and Orfeo ToolBox. *Remote Sens.* **2019**, *11*, 1238. [CrossRef]
77. Heleno, S.; Silveira, M.; Matias, M.; Pina, P. Assessment of supervised methods for mapping rainfall induced landslides in VHR images. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 850–853.
78. Liu, T.; Abd-Elrahman, A. An Object-Based Image Analysis Method for Enhancing Classification of Land Covers Using Fully Convolutional Networks and Multi-View Images of Small Unmanned Aerial System. *Remote Sens.* **2018**, *10*, 457. [CrossRef]
79. Liu, T.; Abd-Elrahman, A.; Morton, J.; Wilhelm, V.L. Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system. *GIScience Remote Sens.* **2018**, *55*, 243–264. [CrossRef]
80. Liu, T.; Abd-Elrahman, A. Deep convolutional neural network training enrichment using multi-view object-based analysis of Unmanned Aerial systems imagery for wetlands classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *139*, 154–170. [CrossRef]
81. Liu, T.; Abd-Elrahman, A.; Dewitt, B.; Smith, S.; Morton, J.; Wilhelm, V.L. Evaluating the potential of multi-view data extraction from small Unmanned Aerial Systems (UASs) for object-based classification for Wetland land covers. *GIScience Remote Sens.* **2019**, *56*, 130–159. [CrossRef]
82. Pande-Chhetri, R.; Abd-Elrahman, A.; Liu, T.; Morton, J.; Wilhelm, V.L. Object-based classification of wetland vegetation using very high-resolution unmanned air system imagery. *Eur. J. Remote Sens.* **2017**, *50*, 564–576. [CrossRef]
83. Liu, T.; Yang, L. A Fully Automatic Method for Rapidly Mapping Impacted Area by Natural Disaster. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, 26 September–2 October 2020; pp. 6906–6909.
84. Liu, T.; Yang, L.; Lunga, D.D. Towards Misregistration-Tolerant Change Detection using Deep Learning Techniques with Object-Based Image Analysis. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA, 5–8 November 2019; pp. 420–423.
85. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
86. Yang, H.; Yu, B.; Luo, J.; Chen, F. Semantic segmentation of high spatial resolution images with deep neural networks. *Gisci. Remote Sens.* **2019**, *56*, 749–768. [CrossRef]
87. Chen, T.; Qiu, C.; Schmitt, M.; Zhu, X.; Sabel, C.; Prishchepov, A. Mapping horizontal and vertical urban densification in Denmark with Landsat time-series from 1985 to 2018: A semantic segmentation solution. *Remote Sens. Environ.* **2020**, *251*, 112096. [CrossRef]
88. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [CrossRef]
89. Wei, P.; Chai, D.; Lin, T.; Tang, C.; Du, M.; Huang, J. Large-scale rice mapping under different years based on time-series Sentinel-1 images using deep semantic segmentation model. *Isprs J. Photogramm. Remote Sens.* **2021**, *174*, 198–214. [CrossRef]
90. Zhang, D.; Pan, Y.; Zhang, J.; Hu, T.; Zhao, J.; Li, N.; Chen, Q. A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution. *Remote Sens. Environ.* **2020**, *247*, 111912. [CrossRef]
91. Zhang, Y.; Chen, G.; Vukomanovic, J.; Singh, K.; Liu, Y.; Holden, S.; Meentemeyer, R. Recurrent Shadow Attention Model (RSAM) for shadow removal in high-resolution urban land-cover mapping. *Remote Sens. Environ.* **2020**, *247*, 111945. [CrossRef]
92. Li, Z.; Shen, H.; Cheng, Q.; Liu, Y.; You, S.; He, Z. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 197–212. [CrossRef]
93. Schiefer, F.; Kattenborn, T.; Frick, A.; Frey, J.; Schall, P.; Koch, B.; Schmidtlein, S. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 205–215. [CrossRef]
94. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
95. Schmitt, M.; Prexl, J.; Ebel, P.; Liebel, L.; Zhu, X.X. Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping–Challenges and Opportunities. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *V-3-2020*, 795–802. [CrossRef]
96. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4981–4990.

97. Vernaza, P.; Chandraker, M. Learning random-walk label propagation for weakly-supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision And pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2953–2961. [CrossRef]
98. Shi, Q.; Liu, X.; Huang, X. An active relearning framework for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 3468–3486. [CrossRef]
99. Robinson, C.; Malkin, K.; Jojic, N.; Chen, H.; Qin, R.; Xiao, C.; Schmitt, M.; Ghamisi, P.; Hänsch, R.; Yokoya, N. Global Land-Cover Mapping With Weak Supervision: Outcome of the 2020 IEEE GRSS Data Fusion Contest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3185–3199. [CrossRef]
100. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 876–885.
101. Li, H.; Wang, Y.; Xiang, S.; Duan, J.; Zhu, F.; Pan, C. A label propagation method using spatial-spectral consistency for hyperspectral image classification. *Int. J. Remote Sens.* **2016**, *37*, 191–211. [CrossRef]
102. Qiao, R.; Ghodsi, A.; Wu, H.; Chang, Y.; Wang, C. Simple weakly supervised deep learning pipeline for detecting individual red-attacked trees in VHR remote sensing images. *Remote Sens. Lett.* **2020**, *11*, 650–658. [CrossRef]
103. Wei, Y.; Ji, S. Scribble-Based Weakly Supervised Deep Learning for Road Surface Extraction From Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5602312. [CrossRef]
104. Carlson, T.N.; Ripley, D.A. On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sens. Environ.* **1997**, *62*, 241–252. [CrossRef]
105. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sens.* **2020**, *12*, 207. [CrossRef]
106. Gao, B.-C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
107. Immitzer, M.; Neuwirth, M.; Böck, S.; Brenner, H.; Vuolo, F.; Atzberger, C. Optimal input features for tree species classification in Central Europe based on multi-temporal Sentinel-2 data. *Remote Sens.* **2019**, *11*, 2599. [CrossRef]
108. Zhang, L.; Ma, J.; Lv, X.; Chen, D. Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 117–121. [CrossRef]
109. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Volume 9351, pp. 234–241.
110. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; Wang, R. Deep few-shot learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2290–2304. [CrossRef]
111. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–34. [CrossRef]
112. Luo, N.; Wan, T.; Hao, H.; Lu, Q. Fusing high-spatial-resolution remotely sensed imagery and OpenStreetMap data for land cover classification over urban areas. *Remote Sens.* **2019**, *11*, 88. [CrossRef]
113. Wan, T.; Lu, H.; Lu, Q.; Luo, N. Classification of high-resolution remote-sensing image using openstreetmap information. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2305–2309. [CrossRef]
114. Comandur, B.; Kak, A.C. Semantic Labeling of Large-Area Geographic Regions Using Multi-View and Multi-Date Satellite Images, and Noisy OSM Training Labels. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4573–4594. [CrossRef]
115. Zhang, R.; Albrecht, C.; Zhang, W.; Cui, X.; Finkler, U.; Kung, D.; Lu, S. Map Generation from Large Scale Incomplete and Inaccurate Data Labels. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 6–10 July 2020; pp. 2514–2522.
116. Wang, S.; Di Tommaso, S.; Faulkner, J.; Friedel, T.; Kennepohl, A.; Strey, R.; Lobell, D.B. Mapping crop types in southeast india with smartphone crowdsourcing and deep learning. *Remote Sens.* **2020**, *12*, 2957. [CrossRef]
117. Sun, B.; Feng, J.; Saenko, K. Return of frustratingly easy domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
118. Pan, S.J.; Kwok, J.T.; Yang, Q. Transfer learning via dimensionality reduction. In Proceedings of the AAAI, Chicago, IL, USA, 13–17 July 2008; pp. 677–682.
119. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [CrossRef]
120. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer feature learning with joint distribution adaptation. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 8–12 April 2013; pp. 2200–2207.
121. Matasci, G.; Volpi, M.; Kanevski, M.; Bruzzone, L.; Tuia, D. Semisupervised transfer component analysis for domain adaptation in remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3550–3564. [CrossRef]
122. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 443–450.
123. Liu, W.; Su, F.; Jin, X.; Li, H.; Qin, R. Bispace Domain Adaptation Network for Remotely Sensed Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2020**. [CrossRef]

124. Tasar, O.; Happy, S.; Tarabalka, Y.; Alliez, P. Colormapgan: Unsupervised domain adaptation for semantic segmentation using color mapping generative adversarial networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7178–7193. [CrossRef]
125. Ji, S.; Wang, D.; Luo, M. Generative Adversarial Network-Based Full-Space Domain Adaptation for Land Cover Classification From Multiple-Source Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3816–3828. [CrossRef]
126. Zou, M.; Zhong, Y. Transfer learning for classification of optical satellite image. *Sens. Imaging* **2018**, *19*, 6. [CrossRef]
127. Gómez-Chova, L.; Tuia, D.; Moser, G.; Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *Proc. IEEE* **2015**, *103*, 1560–1584. [CrossRef]
128. Audebert, N.; Le Saux, B.; Lefèvre, S. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 20–32. [CrossRef]
129. Robinson, C.; Hou, L.; Malkin, K.; Soobitsky, R.; Czawlytko, J.; Dilkina, B.; Jojic, N. Large scale high-resolution land cover mapping with multi-resolution data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12726–12735.
130. Laurin, G.V.; Liesenberg, V.; Chen, Q.; Guerriero, L.; Del Frate, F.; Bartolini, A.; Coomes, D.; Wilebore, B.; Lindsell, J.; Valentini, R. Optical and SAR sensor synergies for forest and land cover mapping in a tropical site in West Africa. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *21*, 7–16. [CrossRef]
131. Huang, X.; Zhang, L. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 721–732. [CrossRef]
132. Zhang, Q.; Qin, R.; Huang, X.; Fang, Y.; Liu, L. Classification of Ultra-High Resolution Orthophotos Combined with DSM Using a Dual Morphological Top Hat Profile. *Remote Sens.* **2015**, *7*, 16422–16440. [CrossRef]
133. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 161–172. [CrossRef]
134. Sonobe, R.; Yamaya, Y.; Tani, H.; Wang, X.; Kobayashi, N.; Mochizuki, K.-I. Crop classification from Sentinel-2-derived vegetation indices using ensemble learning. *J. Appl. Remote Sens.* **2018**, *12*, 026019. [CrossRef]
135. Gerstmann, H.; Möller, M.; Gläßer, C. Optimization of spectral indices and long-term separability analysis for classification of cereal crops using multi-spectral RapidEye imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 115–125. [CrossRef]
136. Settles, B. Active Learning Literature Survey. 2009. Available online: http://digital.library.wisc.edu/1793/60660 (accessed on 15 November 2021).
137. Luo, T.; Kramer, K.; Goldgof, D.B.; Hall, L.O.; Samson, S.; Remsen, A.; Hopkins, T.; Cohn, D. Active learning to recognize multiple types of plankton. *J. Mach. Learn. Res.* **2005**, *6*, 589–613.
138. Tuia, D.; Pasolli, E.; Emery, W.J. Using active learning to adapt remote sensing image classifiers. *Remote Sens. Environ.* **2011**, *115*, 2232–2242. [CrossRef]
139. Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching networks for one shot learning. *arXiv* **2016**, arXiv:1606.04080.
140. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical networks for few-shot learning. In Proceedings of the Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
141. Li, H.; Cui, Z.; Zhu, Z.; Chen, L.; Zhu, J.; Huang, H.; Tao, C. RS-MetaNet: Deep meta metric learning for few-shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6983–6994. [CrossRef]
142. Cheng, G.; Cai, L.; Lang, C.; Yao, X.; Chen, J.; Guo, L.; Han, J. SPNet: Siamese-prototype network for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5608011. [CrossRef]
143. Li, L.; Han, J.; Yao, X.; Cheng, G.; Guo, L. DLA-MatchNet for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7844–7853. [CrossRef]
144. Tingzon, I.; Orden, A.; Go, K.; Sy, S.; Sekara, V.; Weber, I.; Fatehkia, M.; García-Herranz, M.; Kim, D. Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In Proceedings of the AI for Social Good ICML 2019 Workshop, Long Beach, CA, USA, 10–15 June 2019.
145. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
146. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
147. Daumé III, H. Frustratingly easy domain adaptation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 256–263.
148. Deng, X.; Zhu, Y.; Tian, Y.; Newsam, S. Scale Aware Adaptation for Land-Cover Classification in Remote Sensing Imagery. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 5–9 January 2021; pp. 2160–2169.
149. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
150. ISPRS. ISPRS Semantic Labeling Benchmark Dataset. Available online: http://www2.isprs.org/commissions/comm3/wg4/3d-semantic-labeling.html (accessed on 18 April 2018).
151. Christie, G.; Fendley, N.; Wilson, J.; Mukherjee, R. Functional map of the world. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6172–6180.

152. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic segmentation of aerial images with an ensemble of CNSS. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473–480. [CrossRef]
153. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**. [CrossRef]
154. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4037–4058. [CrossRef] [PubMed]
155. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
156. Ericsson, L.; Gouk, H.; Hospedales, T.M. How Well Do Self-Supervised Models Transfer? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 5414–5423.
157. Adrian, J.; Sagan, V.; Maimaitijiang, M. Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 215–235. [CrossRef]
158. Witharana, C.; Bhuiyan, M.; Liljedahl, A.; Kanevskiy, M.; Epstein, H.; Jones, B.; Daanen, R.; Griffin, C.; Kent, K.; Jones, M. Understanding the synergies of deep learning and data fusion of multispectral and panchromatic high resolution commercial satellite imagery for automated ice-wedge polygon detection. *Isprs J. Photogramm. Remote Sens.* **2020**, *170*, 174–191. [CrossRef]
159. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
160. Ma, W.; Shen, J.; Zhu, H.; Zhang, J.; Zhao, J.; Hou, B.; Jiao, L. A Novel Adaptive Hybrid Fusion Network for Multiresolution Remote Sensing Images Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5400617. [CrossRef]
161. Vivone, G.; Dalla Mura, M.; Garzelli, A.; Restaino, R.; Scarpa, G.; Ulfarsson, M.; Alparone, L.; Chanussot, J. A New Benchmark Based on Recent Advances in Multispectral Pansharpening: Revisiting Pansharpening With Classical and Emerging Pansharpening Methods. *Ieee Geosci. Remote Sens. Mag.* **2021**, *9*, 53–81. [CrossRef]
162. Meng, X.; Xiong, Y.; Shao, F.; Shen, H.; Sun, W.; Yang, G.; Yuan, Q.; Fu, R.; Zhang, H. A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and Implementation. *Ieee Geosci. Remote Sens. Mag.* **2021**, *9*, 18–52. [CrossRef]
163. Loncan, L.; De Almeida, L.B.; Bioucas-Dias, J.M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Fabre, S.; Liao, W.; Licciardi, G.A.; Simoes, M. Hyperspectral pansharpening: A review. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 27–46. [CrossRef]
164. Hu, J.; Hong, D.; Zhu, X. MIMA: MAPPER-Induced Manifold Alignment for Semi-Supervised Fusion of Optical Image and Polarimetric SAR Data. *Ieee Trans. Geosci. Remote Sens.* **2019**, *57*, 9025–9040. [CrossRef]
165. Shao, Z.; Cai, J.; Fu, P.; Hu, L.; Liu, T. Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product. *Remote Sens. Environ.* **2019**, *235*, 111425. [CrossRef]
166. Ghamisi, P.; Rasti, B.; Yokoya, N.; Wang, Q.; Hofle, B.; Bruzzone, L.; Bovolo, F.; Chi, M.; Anders, K.; Gloaguen, R.; et al. Multisource and Multitemporal Data Fusion in Remote Sensing A comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 6–39. [CrossRef]
167. Jia, S.; Zhan, Z.; Zhang, M.; Xu, M.; Huang, Q.; Zhou, J.; Jia, X. Multiple Feature-Based Superpixel-Level Decision Fusion for Hyperspectral and LiDAR Data Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1437–1452. [CrossRef]
168. Hoffman-Hall, A.; Loboda, T.V.; Hall, J.V.; Carroll, M.L.; Chen, D. Mapping remote rural settlements at 30 m spatial resolution using geospatial data-fusion. *Remote Sens. Environ.* **2019**, *233*, 111386. [CrossRef]
169. Chen, Y.; Li, C.; Ghamisi, P.; Jia, X.; Gu, Y. Deep Fusion of Remote Sensing Data for Accurate Classification. *Ieee Geosci. Remote Sens. Lett.* **2017**, *14*, 1253–1257. [CrossRef]
170. Cao, R.; Tu, W.; Yang, C.; Li, Q.; Liu, J.; Zhu, J.; Zhang, Q.; Li, Q.; Qiu, G. Deep learning-based remote and social sensing data fusion for urban region function recognition. *Isprs J. Photogramm. Remote Sens.* **2020**, *163*, 82–97. [CrossRef]
171. Zhu, H.; Ma, W.; Li, L.; Jiao, L.; Yang, S.; Hou, B. A Dual–Branch Attention fusion deep network for multiresolution remote–Sensing image classification. *Inf. Fusion* **2020**, *58*, 116–131. [CrossRef]
172. Bergado, J.R.; Persello, C.; Stein, A. Fusenet: End-to-end multispectral vhr image fusion and classification. In Proceedings of the IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2091–2094.
173. Quan, L.; Li, H.; Li, H.; Jiang, W.; Lou, Z.; Chen, L. Two-Stream Dense Feature Fusion Network Based on RGB-D Data for the Real-Time Prediction of Weed Aboveground Fresh Weight in a Field Environment. *Remote Sens.* **2021**, *13*, 2288. [CrossRef]
174. Qin, N.; Hu, X.; Dai, H. Deep fusion of multi-view and multimodal representation of ALS point cloud for 3D terrain scene recognition. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 205–212. [CrossRef]
175. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
176. Lipton, Z.C.; Berkowitz, J.; Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv* **2015**, arXiv:1506.00019.
177. Weikmann, G.; Paris, C.; Bruzzone, L. TimeSen2Crop: A Million Labeled Samples Dataset of Sentinel 2 Image Time Series for Crop-Type Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4699–4708. [CrossRef]
178. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

179. Russwurm, M.; Korner, M. Self-attention for raw optical Satellite Time Series Classification. *Isprs J. Photogramm. Remote Sens.* **2020**, *169*, 421–435. [CrossRef]
180. Xu, J.; Zhu, Y.; Zhong, R.; Lin, Z.; Xu, J.; Jiang, H.; Huang, J.; Li, H.; Lin, T. DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sens. Environ.* **2020**, *247*. [CrossRef]
181. Zhang, G.; Ghamisi, P.; Zhu, X.X. Fusion of heterogeneous earth observation data for the classification of local climate zones. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7623–7642. [CrossRef]
182. Liu, T.; Abd-Elrahman, A. Multi-view object-based classification of wetland land covers using unmanned aircraft system images. *Remote Sens. Environ.* **2018**, *216*, 122–138. [CrossRef]
183. Ahmad, S.K.; Hossain, F.; Eldardiry, H.; Pavelsky, T.M. A fusion approach for water area classification using visible, near infrared and synthetic aperture radar for South Asian conditions. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2471–2480. [CrossRef]
184. Matasci, G.; Longbotham, N.; Pacifici, F.; Kanevski, M.; Tuia, D. Understanding angular effects in VHR imagery and their significance for urban land-cover model portability: A study of two multi-angle in-track image sequences. *ISPRS J. Photogramm. Remote Sens.* **2015**, *107*, 99–111. [CrossRef]
185. Yan, Y.; Deng, L.; Liu, X.; Zhu, L. Application of UAV-Based Multi-angle Hyperspectral Remote Sensing in Fine Vegetation Classification. *Remote Sens.* **2019**, *11*, 2753. [CrossRef]
186. de Colstoun, E.C.B.; Walthall, C.L. Improving global scale land cover classifications with multi-directional POLDER data and a decision tree classifier. *Remote Sens. Environ.* **2006**, *100*, 474–485. [CrossRef]
187. Su, L.; Chopping, M.J.; Rango, A.; Martonchik, J.V.; Peters, D.P. Support vector machines for recognition of semi-arid vegetation types using MISR multi-angle imagery. *Remote Sens. Environ.* **2007**, *107*, 299–311. [CrossRef]
188. Mahtab, A.; Sridhar, V.; Navalgund, R.R. Impact of surface anisotropy on classification accuracy of selected vegetation classes: An evaluation using multidate multiangular MISR data over parts of Madhya Pradesh, India. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 250–258. [CrossRef]
189. Koukal, T.; Atzberger, C.; Schneider, W. Evaluation of semi-empirical BRDF models inverted against multi-angle data from a digital airborne frame camera for enhancing forest type classification. *Remote Sens. Environ.* **2014**, *151*, 27–43. [CrossRef]
190. Longbotham, N.; Chaapel, C.; Bleiler, L.; Padwick, C.; Emery, W.J.; Pacifici, F. Very high resolution multiangle urban classification analysis. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1155–1170. [CrossRef]
191. Huang, X.; Chen, H.; Gong, J. Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 127–141. [CrossRef]
192. Huang, X.; Yang, J.; Li, J.; Wen, D. Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery. *Isprs J. Photogramm. Remote Sens.* **2021**, *175*, 403–415. [CrossRef]
193. Liu, C.; Huang, X.; Zhu, Z.; Chen, H.; Tang, X.; Gong, J. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities. *Remote Sens. Environ.* **2019**, *226*, 51–73. [CrossRef]
194. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9729–9738.