Convolutional Neural Network

Subjects: Mathematics, Applied
Contributor: Aaron Maxwell

Convolutional neural network (CNN)-based deep learning (DL) has a wide variety of applications in the geospatial and remote sensing (RS) sciences, and consequently has been a focus of many recent studies.

Keywords: accuracy assessment ; thematic mapping ; feature extraction ; object detection ; semantic segmentation ; instance segmentation ; deep learning

1. Introduction

This paper is the second and final component in a series in which we explore accuracy assessment methods used in remote sensing (RS) (CNN) classification, focusing on scene classification, object detection, semantic segmentation, and instance segmentation tasks. The studies reviewed rarely reported a complete confusion matrix to describe classification error; and when a confusion matrix was reported, the values for each entry in the table generally did not represent estimation of population properties (i.e., represent a probabilistic sample of the map). Some of these issues are not unique to RS DL studies; similar issues have been noted regarding traditional RS classification accuracy assessment, for example by Foody [1] and Stehman and Foody [2].

Building upon traditional RS accuracy assessment standards and practices, a literature review, and our own experiences, we argue that it is important to revisit accuracy assessment rigor and reporting standards in the context of CNNs. In order to spur the development of community consensus and further discussion, we therefore offer an initial summary of recommendations and best practices for the assessment of DL products to support research, algorithm comparison, and operational mapping. In the Background section, we provide an overview of the types of mapping problems to which CNN-based DL is applicable, review established standards and best practices in remote sensing, and summarize current RS DL accuracy assessment approaches. The Recommendations and Best Practices section outlines accuracy assessment issues related to RS DL classification, focusing on issues related to partitioning the data, assessing generalization, assessment metrics, benchmark datasets, and reporting standards.

2. Recommendations and Best Practices for RS DL Accuracy Assessment

The geographic arrangement of the partitioning of the three sets of reference data, used for training, validation, and testing, is important, and should be carried out in a manner that supports the research questions posed [3] and produces unbiased accuracy estimates. (US). **Figure 1**a illustrates simple random chip partitioning, in which image chips are created for the entire study area extent, shown in gray, and then these image chips are randomly split into the three partitions without any additional geographic stratification. **Figure 1**b illustrates geographically stratified chip partitioning in which chips are partitioned into geographically separate regions. **Figure 1**c illustrates tessellation stratified random sampling, in which the study area is first tessellated into rectangular regions or some other unit of consistent area.

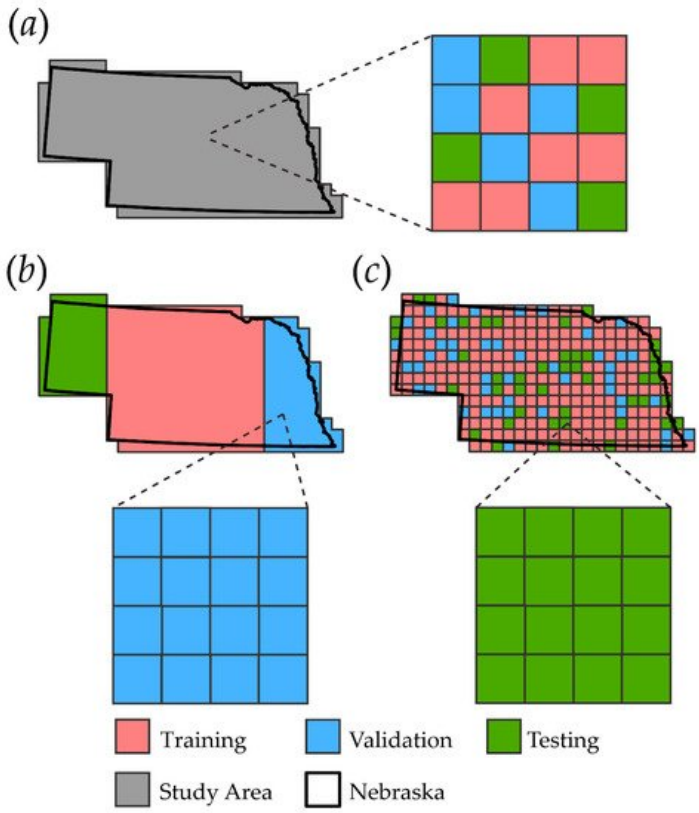

Figure 1. Conceptualization of geospatial sampling options used to generate and partition image chips into training, validation, and testing sets. (**a**) Simple random chip partitioning: entire study area extent is divided into chips, which are then randomly split into training, validation, and testing sets. (**b**) Geographically stratified chip partitioning: stratification using geographically contiguous regions. (**c**) Tessellation stratified random sampling: study area extent is first tessellated and then all chips from the same tessellation unit are subsequently randomly assigned to training, validation, and testing sets.

The choice between the three geographic sampling designs described above has potential consequences for accuracy assessment. Both simple random chip partitioning and tessellation stratified random chip partitioning should yield class proportions for each partition similar to those of the entire study area extent. On the other hand, a potential benefit of both the geographically stratified design and the tessellation stratified random sampling design is that the spatial autocorrelation between the partitions is reduced, and thus the accuracy evaluation may be a more robust test of generalization than, for example, simple random chip partitioning. If so, whatever the geographic sampling design, the subsampling within the partitions should be probabilistic so that unbiased population statistics can be estimated.

In developing training, validation, and testing datasets, it is also important to consider the impact of overlapping image chips. Many studies have noted reduced inference performance near the edge of image chips [4][5][6][7]; therefore, to improve the final map accuracy, it is common to generate image chips that overlap along their edges, and to use only the predictions from the center of each chip in the final, merged output. Generally, the final assessment metrics should be generated from the re-assembled map output, and not the individual chips, since using the chips can result in repeat sampling of the same pixels. For the simple random chip sampling described above, this type of separation may not be possible without subsampling, thus making random chip sampling less attractive.

One of the strengths of many RS DL studies is that they explicitly test generalization within their accuracy assessment design, for example by applying the trained model to new data or new geographic regions not previously seen by the model. Conceptually, model generalization assessment is an extension of the geographically stratified chip partitioning method discussed inSection 3.1, except in this case the new data are not necessarily adjacent to or near the training data, and in many cases they involve multiple replications of different datasets. Assessing model generalization is useful both for adding rigor to the accuracy assessment and for providing insight regarding how the model might perform in an operational environment where new training of the model is impractical every time new data are collected. Examples of such generalization tests include Maggiori et al.

In designing an assessment of model generalization, the type of generalization must be defined, and appropriate testing sets must be developed to address the research question. For example, if the aim is to assess how well a model generalizes to new geographic extents when given comparable data, image chips and reference data in new geographic

regions will be needed. If the goal is to assess generalization to new input data, then it may be appropriate to select the new data from the same geographic extent that was used to train the original model.

As highlighted in Part 1 of this study [8], there are many inconsistencies in the RS DL literature relating to which assessment metrics are calculated, how they are reported, and even the names and terminology used. DL RS studies have primarily adopted measures common in the DL and computer vision communities and have abandoned traditional RS accuracy assessment metrics and terminology. It is important to report metrics that give end users and other researchers insight into how the model and dataset will perform for the application of interest. Below, we recommend best practices for choosing and reporting accuracy assessment metrics for the four CNN-based problem types with these considerations in mind.

In scene classification, the unit of assessment is a single image or chip. Though the image chips are usually generated from geospatial data, the classified images are not normally mapped or referenced back to a specific spatial location in the assessment. In practice, however, it may be challenging to generate a population matrix for scene classification. Many scene classification benchmark datasets appear to be deliberative samples, and usually by design collect samples that are not in proportion to the likelihood of the class in the population.

An example of a scene classification dataset is DeepSat SAT-6 [9] (available from:https://csc.lsu.edu/~saikat/deepsat/; accessed on 1 July 2021). This dataset differentiates six classes (barren, building, grassland, road, tree, and water) and consists of 28-by-28 pixel image chips derived from National Agriculture Imagery Program (NAIP) orthophotography. A total of 405,000 chips are provided, with 324,000 used for training and 81,000 reserved for testing, as defined by the dataset originators. **Figure 2** provides some example image chips included in this dataset.

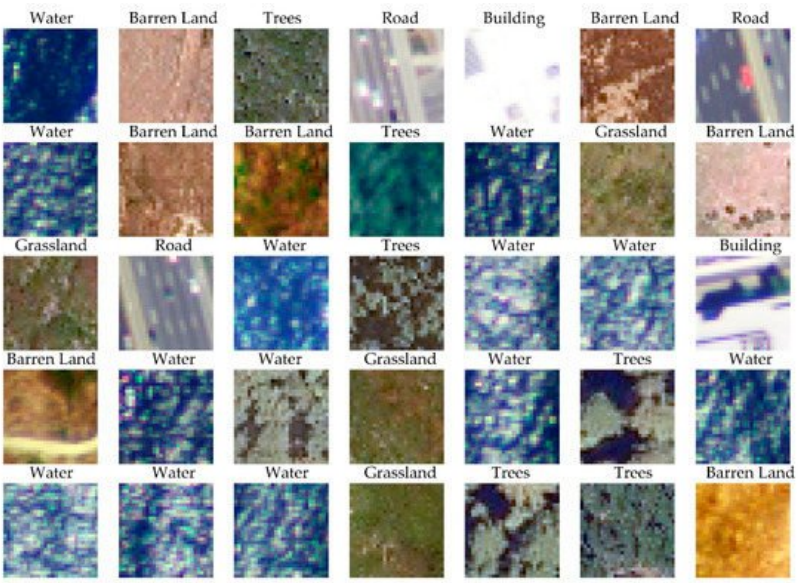

Figure 2. Example DeepSat SAT-6 28-by-28 pixel image chips with associated labels provided by [9].

Table 1 summarizes the accuracy of a ResNet-32 CNN-based scene classification in terms of the confusion matrix with values in the table representing the proportions of the classes in the original testing dataset. Thus, the row and column totals represent the prevalence of each class in the classified data and reference data, respectively. Reporting the entire confusion matrix, along with row and column totals, as well as the class accuracy statistics is useful because it allows greater understanding of how the statistics were calculated, as well as the different components of the classification errors. We use DL terms of precision and recall, as well as the traditional RS terms of UA and PA for clarity.

Table 1. Scene labelling accuracy assessment for DeepSat SAT-6 dataset, with values in the table representing proportions of outcomes based on class prevalence in the benchmark dataset. OA = 0.9967.

		Reference						Row Total	Precision (UA)	F1
		Barren	Building	Grassland	Road	Tree	Water			
Classification	Barren	0.2248	0.0000	0.0006	0.0000	0.0000	0.0000	0.2253	0.9974	0.9943
	Building	0.0000	0.0454	0.0000	0.0000	0.0000	0.0000	0.0455	0.9989	0.9946
	Grassland	0.0020	0.0000	0.1548	0.0000	0.0002	0.0000	0.1570	0.9861	0.9909
	Road	0.0000	0.0004	0.0000	0.0255	0.0000	0.0000	0.0260	0.9819	0.9899
	Tree	0.0000	0.0000	0.0001	0.0000	0.1749	0.0000	0.1750	0.9996	0.9993
	Water	0.0000	0.0000	0.0000	0.0000	0.0000	0.3712	0.3712	1.0000	1.0000
	Column Total	0.2268	0.0459	0.1555	0.0256	0.1751	0.3712			
	Recall (PA)	0.9912	0.9903	0.9957	0.9981	0.9989	1.0000			

Table 2 provides the same data, except this time the columns are normalized to sum to 1.0 (see for example [10]). Because each class has the same column total, this confusion matrix represents a hypothetical case where each class has the same prevalence.

Table 2. Scene labelling accuracy assessment for DeepSat SAT-6 dataset, with class prevalence set to equal the estimated prevalence in the Chesapeake Bay region obtained from [74], and thus the confusion matrix represents an estimate of the population matrix. OA = 0.9978.

		Reference						Row Total	Precision (UA)	F1
		Barren	Building	Grassland	Road	Tree	Water			
Classification	Barren	0.0033	0.0000	0.0010	0.0000	0.0000	0.0000	0.0043	0.7647	0.8633
	Building	0.0000	0.0277	0.0000	0.0000	0.0000	0.0000	0.0277	0.9989	0.9946
	Grassland	0.0000	0.0000	0.2715	0.0000	0.0007	0.0000	0.2721	0.9975	0.9966
	Road	0.0000	0.0003	0.0000	0.0157	0.0000	0.0000	0.0160	0.9803	0.9891
	Tree	0.0000	0.0000	0.0001	0.0000	0.6199	0.0000	0.6200	0.9998	0.9994
	Water	0.0000	0.0000	0.0000	0.0000	0.0000	0.0598	0.0598	1.0000	1.0000
	Column Total	0.0033	0.0280	0.2726	0.0157	0.6206	0.0598			
	Recall (PA)	0.9912	0.9903	0.9957	0.9981	0.9989	1.0000			

Designers of community benchmark datasets, including SAT-6, sometimes do not make it clear whether the proportions of the samples in the various classes represent the prevalence of those classes in the landscape. Thus, it is not clear if **Table 1** is truly a population estimate. However, to illustrate how a population estimate can be obtained from these data, we assumed an application in the East Coast of the USA, and obtained a high spatial resolution (1 m2pixels) map of the 262,358 km2Chesapeake Bay region from [11]. In **Table 2**, the values in the table have been normalized so that the column totals represent the class prevalence values determined from the Chesapeake reference map, and thus, unlike the other two tables, **Table 2** provides an estimate of a potential real-world application of the dataset.

This emphasizes that class prevalence affects most summary accuracy metrics, and therefore the prevalence values used are important. Assuming all classes have equal prevalence, as in **Table 3**, appears to be the standard for RS scene classification. In our survey of 100 papers, five of the 12 studies that dealt with scene classification reported confusion matrices, and all five used this normalization method. However, a hypothetical landscape in which all classes exist in equal proportions is likely to be rare, if it is found at all.

Table 3. Scene labelling accuracy assessment for DeepSat SAT-6 dataset, with accuracy values in each column summing to 1.0, an approach often used in DL studies. OA = 0.9957.

		Reference						Row Total	Precision (UA)	F1
		Barren	Building	Grassland	Road	Tree	Water			
Classification	Barren	0.9912	0.0000	0.0037	0.0000	0.0000	0.0000	0.9949	0.9962	0.9937
	Building	0.0000	0.9903	0.0000	0.0019	0.0000	0.0000	0.9922	0.9981	0.9942
	Grassland	0.0088	0.0000	0.9957	0.0000	0.0011	0.0000	1.0056	0.9902	0.9929
	Road	0.0000	0.0097	0.0002	0.9981	0.0000	0.0000	1.0079	0.9902	0.9941
	Tree	0.0000	0.0000	0.0004	0.0000	0.9989	0.0000	0.9993	0.9996	0.9993
	Water	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000	1.0000
	Column Total	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000			
	Recall (PA)	0.9912	0.9903	0.9957	0.9981	0.9989	1.0000			

As an alternative, some studies only report recall, on the basis that this metric is not affected by prevalence [12][13]. However, **Table 2** highlights the potential hazard of such an approach: the classification has values for recall (PA) above 0.99 for all classes, but the barren class has a precision (UA) value of only 0.76. If only recall values are tabulated, the user would be misled as to the reliability of the classification for barren.

The typical output of a DL semantic segmentation is a wall-to-wall, pixel-level, multi-class classification, similar to that of traditional RS classification. For such classified maps, traditional RS accuracy assessment methods, including using a probabilistic sampling scheme to facilitate unbiased estimates of population-based accuracies, and reporting a complete confusion matrix, has direct application. Using traditional terminology, such as UA and PA, would facilitate communication with RS readers, but almost all authors seem to prioritize communication with the DL community and use the computer science or computer vision terms, such as precision and recall. Given that there are so many of these alternative names for the standard RS metrics, perhaps the most important issue is that all metrics should be clearly defined.

Table 4 gives an example confusion matrix for a semantic segmentation, andTable To produce this table, the LandCover.ai multiclass dataset [14] (available fromhttps://landcover.ai/; accessed on 1 July 2021) was classified with an augmented UNet architecture [15], using the training and testing partitions defined by the originators. In **Table 4**, the values are proportions of the various combinations of reference and classified classes in the landscape, and the column totals represent the class prevalence values. For example, Buildings is a rare class, making up just 0.8% of the reference landscape.

Table 4. Confusion matrix for semantic segmentation of LandCover.ai dataset. Values in the table represent estimates of the population (i.e., landscape) proportions. OA = 95.0%.

		Reference				Row Total	Precision (UA)	F1 Score
		Buildings	Woodlands	Water	Other			
Classification	Buildings	0.007	0.000	0.000	0.003	0.009	0.827	0.716
	Woodlands	0.000	0.325	0.000	0.019	0.344	0.931	0.937
	Water	0.000	0.001	0.053	0.004	0.058	0.961	0.938
	Other	0.001	0.023	0.002	0.562	0.588	0.956	0.955
	Column Total	0.008	0.349	0.056	0.588			
	Recall (PA)	0.705	0.944	0.916	0.955			

Table 4 illustrates why the F1 score on its own, without the values of the constituent precision and recall measures, provides only limited information. The F1 scores of Woodlands and Water are almost identical, suggesting the classification performance is basically the same for these two classes. **Table 4**, however, shows that Water, unlike Woodlands, had a much lower recall than precision.

Surprisingly, although our review of 100 DL papers found the use of many different accuracy metrics, none used some of the more recently developed RS metrics, such as quantity disagreement (QD) and allocation disagreement (AD) suggested by Pontius and Millones [16], which provide useful information regarding the different components of error. For

the map summarized in **Table 4**, the QD was 0.0% and the AD was 5.0%, indicating almost no error is derived from an incorrect estimation of the proportion of the classes; instead, the error derives from the mislabeling of pixels.

Table 5 was derived from [5], which used semantic segmentation to extract the extent of surface mining from historic topographic maps. The accuracy assessment included a geographic generalization component, and was carried out on the mosaicked images, not individual image chips. However, variability in the accuracy of mapping the mining classes and issues of FPs and FNs were captured by precision, recall, and the F1 score. Reporting only the F1 score would obscure the fact that in Geographic Region 4, the much lower accuracy is due to a low recall, whereas the precision is similar to that of the other areas.

Table 5. Binary classification accuracy assessment from prior study [68] in which mining features were extracted from historic topographic maps. NPV = Negative Predictive Value.

Geographic Region	Precision	Recall	Specificity	NPV	F1 Score	OA
1	**0.914**	**0.938**	**0.999**	**0.993**	**0.919**	**0.999**
2	**0.883**	**0.915**	**0.999**	**0.999**	**0.896**	**0.998**
3	**0.905**	**0.811**	**0.998**	**0.993**	**0.837**	**0.992**
4	**0.910**	**0.683**	**0.998**	**9.983**	**0.761**	**0.983**

CNN-based deep learning can also be used to generate spatial probabilistic models. [17] explored DL for the probabilistic prediction of severe hailstorms, while Thi Ngo et al. If the primary output will be a probabilistic model, probabilistic-based assessment metrics should be reported. Thus, precision (i.e., UA) is not assessed, which can be especially misleading when class proportions are imbalanced.

Object detection and instance segmentation both identify individual objects. Since the number of true negative objects (i.e., objects not part of the category of interest) is generally not defined, reporting a full confusion matrix and the associated OA metric is generally not possible. Nevertheless, reporting the number of TP, TN, FN, and FP, along with the derived accuracy measures typically used—precision, recall, and F1 (e.g., [18])—ensures clarity in the results.

A further complication for object detection is that there is normally a range of confidence levels in the detection of any single object. In the context of object detection, however, the AUC PR is normally referred to as average precision (AP), and/or mean average precision (mAP). The second component of object probability, the delineation of the object, is usually quantified in terms of the IoU of the reference and predicted masks or bounding boxes. However, because the choice of a threshold is usually arbitrary, a range of thresholds may be chosen, for example, from 0.50 to 0.95, with steps of 0.05, which would generate 10 sets of object detections, each with its own P-R curve and AP metric.

Unfortunately, there is considerable confusion regarding the AP/mAP terminology. First, these terms may be confused with the average of the precision values in a multiclass classification (e.g. a semantic segmentation). Second, mAP is generally used to indicate a composite AP, for example, typically averaged over different classes [19][20][21], but also sometimes different IoU values [22]. However, because of inconsistencies in the usage of the terms AP and mAP in the past, some sources (e.g., [23]) no longer recommend differentiating between them, and instead use the terms interchangeably.

Because of the lack of standardization in AP and mAP terminology, it is important for accuracy assessments to clearly report how these accuracy metrics were calculated, for example, specifying the IoU thresholds used. Adding subscripts for IoU thresholds, if thresholds are used, or BB or M to indicate whether the metrics are based on bounding boxes or pixel-level masks, can be an effective way to communicate this information (for example, APBBor IoUM). [24] used both superscripts and subscripts to differentiate six AP metrics derived from bounding box and pixel-level masks, as well as three different sizes of objects (small, medium, and large).

In reporting object-level data, it is important to consider whether a count of features or the area of features is of interest. For example, a model may be used to predict the number of individual trees in an image extent. In such a case, it would be appropriate to use the numbers of individuals as the measurement unit. However, if the aim is to estimate the area covered by tree canopy, then it would be more appropriate to use assessment metrics that incorporate the relative area of the bounding boxes or masks.

A variety of benchmark geospatial datasets are available to test new methods and compare models. [25] provide an extensive comparison of benchmark datasets. Despite the range of publicly available benchmark datasets, there are notable gaps in the available data. and/or data are limited, and there is also generally a lack of non-image benchmark datasets, such as those representing LiDAR point clouds, historic cartographic maps, and digital terrain data, as noted in our prior studies [5][6].

Benchmark datasets and the associated metadata inherently constrain the potential for rigorous accuracy assessment, depending on the nature of the data and, crucially, how they are documented. Benchmark datasets in many cases provide pre-determined protocols for accuracy assessment. Therefore, it is particularly helpful if benchmark datasets support best practices, as described below: If the dataset is meant to support the assessment of model generalization to new data and/or geographic extents, the recommended validation configuration should be documented.

In order to improve consistency between studies, replication capacity, and inference potential, it is important to report, in detail, the research methods and accuracy assessment metrics. Here, we provide a list of reporting standards and considerations that will promote clarity for interpretation and replication of experiments.

3. Outstanding Issues and Challenges

First, comparison of new methods and architectures is hindered by the complexity of hyperparameter optimization and architecture design [26][27][28][29]. However, due to lengthy training times, computational demands, and a wide variety of hyperparameters and architecture changes that can be assessed, such systematic experimentation is currently not possible for CNN-based DL [26][27][28][29][30]. When new algorithms or methods are proposed, it is common to use existing machine learning or DL methods as a benchmark for comparison; however, we argue that such comparisons are limited due to an inability to optimize the benchmark model fully [31][32]. [32] suggest that reported improvements when using new algorithms and architectures are often overly optimistic due to inconsistencies in the input data, training process, and experimental designs and because the algorithm to which the comparison is being made was not fully optimized.

This issue is further confounded by the lack of a theoretical understanding of the data abstractions and generalizations modeled by CNNs [33]. For example, textural measures, such as those derived from the gray level co-occurrence matrix (GLCM) after Haralick [34], have shown varying levels of usefulness for different mapping tasks; however, one determinant of the value of these measures is whether or not the classes of interest have unique spatial context or textural characteristics that allow for greater separation [35][36][37][38]. Building on existing literature, a better understanding of how CNNs represent textural and spatial context information at different scales could potentially improve our general understanding of how best to represent and model textural characteristics to enhance a wide range of mapping tasks. Furthermore, improved understanding could guide the development of new DL methods and architectures.

Accuracy estimates are usually empirical estimates of map properties, and as with all statistical estimates, they have uncertainties. , we suggest that confidence intervals be reported where possible. For example, confidence intervals can be estimated for overall accuracy [39] and AUC ROC [40] or multiple model runs can be averaged to assess for variability. Such information can be informative, especially when only small differences in assessment metrics between models are observed.

For example, accuracy assessment methods generally assume that all pixels will map perfectly to the defined classes and that feature boundaries are well-defined. The impact of landscape complexity and heterogeneity should also be considered when designing accuracy assessment protocols and interpreting results. For example, heterogenous landscapes may prove more difficult to map in comparison to more homogeneous landscapes, resulting from more boundaries, edges, class transitions, and, potentially, mixed pixels or gradational boundaries [41][42]. Accounting for mixed pixels, gradational boundaries, difficult to define thematic classes, and the defined minimal mapping unit (MMU) are complexities that exist for all thematic mapping tasks, including those relying on DL, which highlight the need for detailed reporting of sample collection and accuracy assessment methods [8][1][2][43][41][44][45][46][47][48].

Such findings can inform DL researchers as to existing mapping difficulties and research needs, which classes are most difficult to differentiate, and which classes are most important to map for different use cases. We argue that considering the existing literature, methods, and complexities relating to specific mapping tasks or landscape types can offer guidance for knowledge gaps and research needs to which DL may be applicable. Similarly, the current body of knowledge offers insights for extracting information from specific data types. For example, extracting information from true color or color infrared aerial imagery with high spatial resolution, limited spectral resolution, and variability in illuminating conditions between images is an active area of research in the field [14][49][50][51][52][53].

References

1. Foody, G.M. Status of land cover classification accuracy assessment. Remote Sens. Environ. 2002, 80, 185–201.
2. Stehman, S.V.; Foody, G.M. Key issues in rigorous accuracy assessment of land cover products. Remote Sens. Environ. 2019, 231, 111199.
3. Stehman, S.V.; Foody, G.M. Others Accuracy assessment. In The SAGE Handbook of Remote Sensing; Sage: London, UK, 2009; pp. 297–309.
4. Zhang, W.; Liljedahl, A.K.; Kanevskiy, M.; Epstein, H.E.; Jones, B.M.; Jorgenson, M.T.; Kent, K. Transferability of the Deep Learning Mask R-CNN Model for Automated Mapping of Ice-Wedge Polygons in High-Resolution Satellite and UAV Images. Remote Sens. 2020, 12, 1085.
5. Maxwell, A.E.; Bester, M.S.; Guillen, L.A.; Ramezan, C.A.; Carpinello, D.J.; Fan, Y.; Hartley, F.M.; Maynard, S.M.; Pyron, J.L. Semantic Segmentation Deep Learning for Extracting Surface Mine Extents from Historic Topographic Maps. Remote Sens. 2020, 12, 4145.
6. Maxwell, A.E.; Pourmohammadi, P.; Poyner, J.D. Mapping the Topographic Features of Mining-Related Valley Fills Using Mask R-CNN Deep Learning and Digital Elevation Data. Remote Sens. 2020, 12, 547.
7. Zhang, W.; Witharana, C.; Liljedahl, A.K.; Kanevskiy, M. Deep Convolutional Neural Networks for Automated Characterization of Arctic Ice-Wedge Polygons in Very High Spatial Resolution Aerial Imagery. Remote Sens. 2018, 10, 1487.
8. Maxwell, A.; Warner, T.; Guillén, L. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review. Remote Sens. 2021, 13, 2450.
9. Basu, S.; Ganguly, S.; Mukhopadhyay, S.; DiBiano, R.; Karki, M.; Nemani, R. DeepSat: A Learning Framework for Satellite Imagery. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York, NY, USA, 3 November 2015; pp. 1–10.
10. Qi, K.; Yang, C.; Hu, C.; Guan, Q.; Tian, W.; Shen, S.; Peng, F. Polycentric Circle Pooling in Deep Convolutional Networks for High-Resolution Remote Sensing Image Recognition. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2020, 13, 632–641.
11. Land Cover Data Project 2013/2014. Available online: (accessed on 29 April 2021).
12. Tharwat, A. Classification assessment methods. Appl. Comput. Inform. 2021, 17, 168–192.
13. Prakash, N.; Manconi, A.; Loew, S. Mapping Landslides on EO Data: Performance of Deep Learning Models vs. Traditional Machine Learning Models. Remote Sens. 2020, 12, 346.
14. Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Zambrzycka, A.; Dziedzic, T. LandCover. ai: Dataset for Automatic Mapping of Buildings, Woodlands and Water from Aerial Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, Nashville, TN, USA, 19–25 June 2021.
15. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv 2015, arXiv:1505.04597.
16. Pontius, R.G., Jr.; Millones, M. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment. Int. J. Remote Sens. 2011, 32, 4407–4429.
17. Ii, D.J.G.; Haupt, S.E.; Nychka, D.W.; Thompson, G. Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. Mon. Weather Rev. 2019, 147, 2827–2845.
18. Pham, M.-T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-Fine: One-Stage Detector of Small Objects under Various Backgrounds in Remote Sensing Images. Remote Sens. 2020, 12, 2501.
19. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-Scale Spatial and Channel-wise Attention for Improving Object Detection in Remote Sensing Imagery. IEEE Geosci. Remote Sens. Lett. 2020, 17, 681–685.
20. Oh, S.; Chang, A.; Ashapure, A.; Jung, J.; Dube, N.; Maeda, M.; Gonzalez, D.; Landivar, J. Plant Counting of Cotton from UAS Imagery Using Deep Learning-Based Object Detection Framework. Remote Sens. 2020, 12, 2981.
21. Henderson, P.; Ferrari, V. End-to-End Training of Object Class Detectors for Mean Average Precision. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 198–213.
22. Zheng, Z.; Zhong, Y.; Ma, A.; Han, X.; Zhao, J.; Liu, Y.; Zhang, L. HyNet: Hyper-scale object detection network framework for multiple spatial resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 2020, 166, 1–14.
23. COCO-Common Objects in Context. Available online: (accessed on 3 April 2021).

24. Wu, T.; Hu, Y.; Peng, L.; Chen, R. Improved Anchor-Free Instance Segmentation for Building Extraction from High-Resolution Remote Sensing Images. Remote Sens. 2020, 12, 2910.
25. 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.
26. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. IEEE Geosci. Remote Sens. Mag. 2017, 5, 8–36.
27. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. IEEE Geosci. Remote Sens. Mag. 2016, 4, 22–40.
28. 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.
29. 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.
30. Koutsoukas, A.; Monaghan, K.J.; Li, X.; Huan, J. Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J. Cheminform. 2017, 9, 1–13.
31. 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 Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12726–12735.
32. Musgrave, K.; Belongie, S.; Lim, S.-N. A Metric Learning Reality Check. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
33. Sejnowski, T.J. The unreasonable effectiveness of deep learning in artificial intelligence. Proc. Natl. Acad. Sci. USA 2020, 117, 30033–30038.
34. Haralick, R.M. Statistical and structural approaches to texture. Proc. IEEE 1979, 67, 786–804.
35. Warner, T. Kernel-Based Texture in Remote Sensing Image Classification. Geogr. Compass 2011, 5, 781–798.
36. Kim, M.; Madden, M.; Warner, T.A. Forest Type Mapping using Object-specific Texture Measures from Multispectral Ikonos Imagery. Photogramm. Eng. Remote Sens. 2009, 75, 819–829.
37. Kim, M.; Warner, T.A.; Madden, M.; Atkinson, D.S. Multi-scale GEOBIA with very high spatial resolution digital aerial imagery: Scale, texture and image objects. Int. J. Remote Sens. 2011, 32, 2825–2850.
38. Fern, C.; Warner, T.A. Scale and Texture in Digital Image Classification. Photogramm. Eng. Remote Sens. 2002, 68, 51–63.
39. Foody, G.M. Sample size determination for image classification accuracy assessment and comparison. Int. J. Remote Sens. 2009, 30, 5273–5291.
40. Cortes, C.; Mohri, M. Confidence Intervals for the Area Under the ROC Curve. Adv. Neural Inf. Process. Syst. 2005, 17, 305–312.
41. Foody, G.M. Harshness in image classification accuracy assessment. Int. J. Remote Sens. 2008, 29, 3137–3158.
42. Fuller, R.; Groom, G.; Jones, A. The Land-Cover Map of Great Britain: An Automated Classification of Landsat Thematic Mapper Data. Photogramm. Eng. Remote Sens. 1994, 60, 553–562.
43. Foody, G. Thematic Map Comparison. Photogramm. Eng. Remote Sens. 2004, 70, 627–633.
44. Stehman, S.V.; Czaplewski, R.L. Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. Remote Sens. Environ. 1998, 64, 331–344.
45. Stehman, S.V.; Wickham, J.D. Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. Remote Sens. Environ. 2011, 115, 3044–3055.
46. Maxwell, A.E.; Warner, T.A. Thematic Classification Accuracy Assessment with Inherently Uncertain Boundaries: An Argument for Center-Weighted Accuracy Assessment Metrics. Remote Sens. 2020, 12, 1905.
47. Foody, G. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. Int. J. Remote Sens. 1996, 17, 1317–1340.
48. Foody, G.M. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. Int. J. Remote Sens. 2005, 26, 1217–1228.
49. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing

Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.

50. Li, X.; Shao, G. Object-Based Land-Cover Mapping with High Resolution Aerial Photography at a County Scale in Midwestern USA. Remote Sens. 2014, 6, 11372–11390.
51. Witharana, C.; Bhuiyan, A.E.; Liljedahl, A.K.; Kanevskiy, M.; Epstein, H.E.; Jones, B.M.; Daanen, R.; Griffin, C.G.; Kent, K.; Jones, M.K.W. 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.
52. Mou, L.; Hua, Y.; Zhu, X.X. Relation Matters: Relational Context-Aware Fully Convolutional Network for Semantic Segmentation of High-Resolution Aerial Images. IEEE Trans. Geosci. Remote Sens. 2020, 58, 7557–7569.
53. Luo, S.; Li, H.; Shen, H. Deeply supervised convolutional neural network for shadow detection based on a novel aerial shadow imagery dataset. ISPRS J. Photogramm. Remote Sens. 2020, 167, 443–457.
