Current Forestry Reports (2023) 9:149–170
https://doi.org/10.1007/s40725-023-00184-3

REMOTE SENSING (J SUAREZ, SECTION EDITOR)

A Systematic Review of Individual Tree Crown Detection and Delineation with Convolutional Neural Networks (CNN)

Haotian Zhao[1] · Justin Morgenroth[1] · Grant Pearse[2] · Jan Schindler[3]

Accepted: 20 March 2023 / Published online: 5 April 2023

Abstract

Purpose of Review Crown detection and measurement at the individual tree level provide detailed information for accurate forest management. To efficiently acquire such information, approaches to conduct individual tree detection and crown delineation (ITDCD) using remotely sensed data have been proposed. In recent years, deep learning, specifically convolutional neural networks (CNN), has shown potential in this field. This article provides a systematic review of the studies that used CNN for ITDCD and identifies major trends and research gaps across six perspectives: accuracy assessment methods, data types, platforms and resolutions, forest environments, CNN models, and training strategies and techniques.

Recent Findings CNN models were mostly applied to high-resolution red–green–blue (RGB) images. When compared with other state-of-the-art approaches, CNN models showed significant improvements in accuracy. One study reported an increase in detection accuracy of over 11%, while two studies reported increases in F1-score of over 16%. However, model performance varied across different forest environments and data types. Several factors including data scarcity, model selection, and training approaches affected ITDCD results.

Summary Future studies could (1) explore data fusion approaches to take advantage of the characteristics of different types of remote sensing data, (2) further improve data efficiency with customised sample approaches and synthetic samples, (3) explore the potential of smaller CNN models and compare their learning efficiency with commonly used models, and (4) evaluate impacts of pre-training and parameter tunings.

Keywords Deep learning · Tree detection · Crown delineation · Forestry · Remote sensing · Object detection · Instance segmentation

Introduction

Effective management of trees is reliant upon accurate and up-to-date information, including a description of individual crowns canopy cover (CC). Canopy cover is a basic metric describing the horizontal spread and distribution of tree crowns; it is typically expressed as the percentage of total ground area covered by a 2D projection of tree canopy. The ease of determining CC and its conceptual simplicity leads to CC being a key goal or target in many forest or tree plans, including in the United Nations' State of the World's Forests technical reports [1]. Despite its ubiquity, CC mapping fails to distinguish between individual tree crowns. This is a limitation as CC mapping cannot be used to effectively describe other important characteristics such as variation in crown sizes, individual tree location, species, or the health of individual trees. This additional level of detail is important for the analysis, modelling, or management, of a variety of environments, including natural forests, planted forests, urban forests, and orchards. Individual tree crown mapping has been found to support more accurate analysis of carbon storage estimation [2], biodiversity assessment [3], urban forest management [4, 5], canopy closure estimation [6], ecosystem service modelling [7], and forest health description [8]. In summary, without being able to uniquely identify individual tree crowns from canopy cover maps, we can only

✉ Haotian Zhao
hzh159@uclive.ac.nz

[1] New Zealand School of Forestry, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

[2] Data and Geospatial Intelligence, Scion, 49 Sala Street, Private Bag 3020, Rotorua 3046, New Zealand

[3] Informatics, Manaaki Whenua – Landcare Research, PO Box 10345, The Terrace, Wellington 6143, New Zealand

provide a basic overview of the horizontal structure of treed environments.

There are various ways to map individual tree crowns. Field measurement of individual tree crowns can be a time-consuming process, especially when dealing with a large number of trees across large areas. As a result, ground-based assessments are typically used for small survey areas or small numbers of sample plots. Access issues (privately-owned or otherwise inaccessible land) can limit the ability to measure tree crowns in the field [9, 10]. Compared with field surveys, remote sensing can provide tree cover information at a large scale, more cost-effectively, and is not limited by access issues. A conventional way of extracting individual crown attributes from remotely sensed data sources (e.g. aerial imagery, lidar, stereo images) is by manually delineating trees based on a visual assessment [11–13]. Like fieldwork, manual delineation is labour-intensive and subjective. In the past decade, numerous methods have been developed to automate individual tree detection and crown delineation (ITDCD) from remote sensing data. These methods are mostly based on image analysis techniques or, more recently, deep learning approaches [14••].

Image analysis techniques have been developed and applied for ITDCD since the 1990s [15]. In general, these techniques can be divided into two categories, individual tree detection and individual crown delineation. The former extracts tree locations or treetops as point features, while the latter delineates individual crowns as polygon areas. Some commonly known individual tree detection methods include local maximum filtering, image binarisation, template matching, and scale analysis. Likewise, for individual crown delineation, some widely used methods include valley-following, region-growing, and watershed segmentation [16]. For some crown delineation methods, individual tree detection is required as a preliminary step.

Although many improvements and developments have been made to image analysis techniques, there is no consensus on which methods are optimal for use with different image types and forest conditions. For applications under similar scenarios, methodological comparisons may also be difficult due to the varied approaches used in model evaluation. However, a common principle of those techniques is the topographic analogy of the canopy surface, which assumes the crown centres have a higher reflectance or elevation than other parts of the same crown. This assumption can be limiting when those techniques are applied to forests dominated by broadleaf species. This is because image analysis techniques generally produce better results on conifers due to their excurrent form and apical dominance, rather than on broadleaf species with their decurrent form with multiple leaders [17, 18]. Some studies have tried to improve the performance of methods like local maximum filtering, watershed, and region-growing by applying multiscale analysis [19, 20] or by integrating morphological features [21, 22], but generalising those models to different scenes is time-consuming as it requires further analysis on forest features and manual adjustments on parameters.

Deep learning (DL), a subclass of machine learning, has developed rapidly in the past decade, benefiting from improved computing abilities [23]. As a type of neural network, DL models consist of layers of interconnected processing nodes, or neurons, to simulate the structure and functions of the human brain. When interpreting data, the early or shallow layers extract low-level features and pass these to the deeper or later layers, which extract high-level, abstract features. With more connected layers, the network becomes deeper and gains greater abilities to comprehensively represent data. Unlike machine learning models, which require data transformations or feature selections before conducting subtasks (e.g. object-based image analysis), DL takes data in its raw form and automatically identifies features used for detection or classification [24••]. When training a DL model, data are passed through the network multiple times. The internal parameters that control the activations of neurons are adjusted based on desired outputs. Due to the complex connections of neurons, such processes usually require a large amount of data and are computationally expensive. However, since DL networks are parallelisable, DL has benefited from advancements of graphics processing units (GPU), which can use thousands of cores to distribute calculations and speed up the training process.

In image analysis tasks, specifically object detection and instance segmentation, a commonly used DL model is the convolutional neural network (CNN). CNN adds convolutional layers to the fully connected neural network, which extract local features at different levels by sliding multiple window filters across an image. Each filter is a matrix of weights and has a much smaller size than the input image to allow extractions of small, meaningful features, such as edges, from a location [25]. These convolutional operations take both pixel values and the local arrangement of features into consideration and therefore can extract more distinctive representations from an image [23].

Since 2012, CNN-based models have become state-of-the-art in many image-analysis challenges and have been applied to remote sensing tasks, including ITDCD. CNN models have achieved high accuracies (over 90%) in both individual tree detection [26, 27] and crown delineation [28, 29]. In some studies where CNN models were compared with traditional techniques for ITDCD using remote sensing imagery, significant improvements in both accuracy and inference speed were observed [14••, 30••].

Compared with image analysis techniques, using CNN models for ITDCD has two major advantages. Firstly, CNN models have transfer-learning ability, such that knowledge learned from one site can be transferred to another site,

resulting in model generalisability. In contrast, most image analysis models rely on pre-defined parameters, which need to be manually adjusted when applied to a different environment or dataset. Secondly, CNN models detect trees at an object level by learning patterns from hierarchical combinations of image features [30••]. While other techniques describe tree objects solely based on reflectance or morphological features, CNN models provide a more comprehensive representation of tree appearance with abstract visual characteristics [28].

In recent years, there have been an increasing number of studies applying CNN models to ITDCD tasks with different data and in various forest environments. There have even been a small number of reviews that are generally related to the topic. These include DL for earth observation [23, 31], CNNs for vegetation analysis [24••], and other general reviews for DL use in remote sensing [32, 33]. Diez et al. [34] reviewed DL forestry applications that specifically used red-green-blue (RGB) imagery captured by unmanned autonomous vehicles (UAVs); five ITDCD-relevant studies published between 2017 and March 2021 were included, but the methods used by those studies were not compared or discussed in detail. Thus, while there have been some reviews on the use of CNNs in forested environments, a key gap remains. The present review will address this gap by performing a systematic review of the literature on the use of CNN for ITDCD across six key themes including accuracy assessment methods, data types, platforms and resolutions, forest environments, CNN models, and training strategies and techniques. The results of this systematic review will provide a comprehensive summary of relevant studies, give researchers a wider understanding of the key findings and limitations of the latest approaches, and identify possible directions for future research.

Systematic Literature Review Methods

Search Strategy

The systematic review was undertaken using a framework called "preferred reporting items for systematic reviews and meta-analyses" (PRISMA) [35]. The Scopus database was queried using the string: ("convolutional neural network" OR "convolutional network" OR "convolution neural network" OR "deep convolution network" OR *cnn* OR convnet OR "deep learning") AND (tree OR canop* OR crown* OR forest* OR plant*). The search fields were limited to abstract, keywords, and title, while the document type was limited to published journal articles with full access in English or Chinese. The search was further constrained to include only articles published between January 2012, the year during which deep learning approaches began to overtake other machine learning methods and win some well-known computer vision competitions [31, 36], and December 2021.

Article Selections

After the initial query, search results ($n = 7493$ results) were screened by subject area and title to exclude articles that were not related to the topic, for example, medical image analysis or studies that focused on tree-like structures within CNN algorithms. This reduced the number of articles to 602. After further reviewing the article keywords and abstracts for relevance, 103 articles were retained. Each of these articles was checked for inclusion eligibility by reviewing their complete contents. The criteria for inclusion were:

- The study used CNN-based deep learning methods for individual tree detection or crown delineation.
- The study used remote sensing data, specifically including RGB, multispectral or lidar data captured from UAV, aerial, or satellite platforms. Studies using street-level images or terrestrial lidar were excluded.
- The study included a formal means of assessing the accuracy of tree crown detection or delineation. Specifically, this meant assessment of the extent of a bounding box for tree detection or assessment of the mask area or bounding polygon in the case of crown delineation. Studies using bounding boxes for tree counting without an assessment of crown extents were excluded.

Based on these criteria, 35 articles were ultimately included in this review. A diagram of the selection process is shown in the Appendix (Fig. 3).

The included articles were selected from 20 different journals, with the journal *Remote Sensing* containing the most studies (11 out of 35). The earliest published year is 2019 with only 4 articles. This number later increased to 13 in 2020 and 14 in 2021. This indicates that applying CNN to ITDCD tasks is a growing area of research, especially when compared with ITDCD using other image analysis techniques that have been well-studied for decades. The articles and journal information can be found in the Appendix (Table 7).

Data Synthesis

The information from the articles were extracted from multiple perspectives, including ITDCD task, extracted crown features, dataset, training process, accuracy assessment, and outputs. Information summarised from those perspectives were then synthesised and discussed in the context of six review themes. A summary of synthesis items and descriptions is shown in the Appendix (Table 8).

Review of Key Themes in the Literature

Accuracy Assessment Methods

ITDCD accuracy assessment includes both qualitative and quantitative evaluations of mapped tree locations and crown delineations. For qualitative evaluation, visual inspections are usually conducted to compare the mapping results with the reference data such as aerial imagery or a canopy height model (CHM). Visual inspection not only provides a direct sense of the performance of ITDCD methods but also a quick overview of the spatial distribution of errors [17]. For quantitative evaluations, ITDCD methods are most often evaluated using metrics commonly used in computer vision. These metrics reflect the performances of individual tree detection or individual crown delineation from various perspectives. Using combinations of these metrics can provide a comprehensive understanding of the ITDCD results and sources of error. These metrics also provide users with awareness of a model's performance so that it can be used and interpreted appropriately in the context of forest planning and management. Detailed descriptions of those metrics are shown in Table 1.

Precision and recall (PR; $n = 34$), F1-score ($n = 21$), and average precision (AP; $n = 15$) were the most used metrics among reviewed literature. Precision reflects whether tree crowns are detected correctly, with few false positives. Recall reflects a model's ability to find most trees from the imagery, implying the model does not fail to detect many reference trees. Both F1-score and AP provide a unified measurement of PR, but the two metrics are fundamentally different. F1-score tends to reflect the balance between PR by calculating harmonic means. It represents a certain point on the precision and recall curve. In contrast, AP estimates a model's performance by calculating an average of multiple precision values across recall points. It represents an area under the precision and recall curve. A larger area reflects a higher accuracy. In this review, two studies [38, 39] included both F1-score and AP to provide a more comprehensive evaluation of their models.

Apart from PR, F1-score, and AP, other metrics were also applied to evaluate ITDCD results in specific situations. The mAP, which averages the AP across all detection classes, was used for multiclass ITDCD [40, 41]. Overall accuracy [27, 29] and detection percentage [42] were applied to assess

Table 1 Accuracy metrics identified from reviewed studies and their descriptions

Metric name	Description
Intersection over union (IoU)	The area of overlap between ground-truth and detected results (either as a bounding box or polygon mask) divided by their combined area. IoU is usually used to define a true detection $\text{IoU} = \frac{\text{Predicted} \cap \text{Ground Truth}}{\text{Predicted} \cup \text{Ground Truth}}$
True positive (TP)	The number of correctly detected trees
False negative (FP)	The number of missing reference trees in detection results
False positive (FP)	The number of other objects that have been mis-detected as trees
Precision	The percentage of correctly detected crowns among all the objects that the model has identified. A high precision indicates the detected results contain a high proportion of correctly detected trees and few false positives $\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$
Recall	The percentage of correctly detected trees among all reference trees. A high recall indicates the model is able to find most trees from the imagery and does not fail to detect many reference trees $\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$
Average precision (AP)	The average value of multiple precision values at different recall levels from 0 to 1 with a certain number of steps [37]. It represents the area below the precision-recall curve. AP reflects the trade-off between precision and recall and is a widely used *v* for evaluating the performance of object detection models $AP = \sum_{k=0}^{k=n-1} [\text{Recall}(k) - \text{Recall}(k+1)]\text{Precision}(k)$ $\text{Recall}(n) = 0, \text{Precision}(n) = 1$ $n = \text{Number of steps.}$
Mean average precision (mAP)	The average value of AP across different target classes $mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k$ $AP_k = the\ AP \text{ of class } k$ $n = \text{the number of target classes}$
F1-score	F1-score combines both precision and recall and provides a comprehensive assessment of a model. It is defined as the harmonic mean of recall and precision. A higher F1-score indicates a better balance between recall and precision $\text{F1} - \text{score} = 2\frac{\text{Precision Recall}}{\text{Precision} + \text{Recall}}$
Overall accuracy	$\text{OA} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative} + \text{False Positive}}$

the correct detection ratio. RMSE and R^2 were used to evaluate the distribution of tree count numbers [29].

The inconsistencies between studies exist not only in the use of accuracy metrics but also in the way those metrics are defined. For example, Chadwick et al. [28] defined AP as the average values of precision across image tiles rather than calculating it across recall levels. Intersection over union (IoU) was also set at multiple levels in some studies in contrast to a common choice of 50%. When choosing a higher IoU threshold, a true positive prediction, either a bounding box or a segmented mask, is required to overlap more accurately with the ground truth area. Introducing multiple IoU thresholds will result in different ways of calculating metrics including AP and mAP but can provide a more detailed view of the quality of localisation. In the reviewed studies, Ammar et al. [40] calculated AP, mAP, and F1-score at five IoU levels ranging from 50 to 90%. Xi et al. [43] and Chiang et al. [38] calculated AP over a range of IoUs between 50 and 95%, which aligns with the standard defined in common objects in context (COCO), a widely used dataset for computer vision competition. It should be noted that this cross-IoU AP is defined as mAP in COCO [44, 45].

Comparing CNN models with traditional ITDCD approaches across studies is difficult. One reason is that CNN introduces the concept of IoU and defines the TP in a different way from traditional ITDCD studies. Besides, CNN tends to use more RGB images, while traditional ITDCD applies more to lidar or fusion datasets [18, 46]. However, several studies have shown a significant advantage of CNN models. Braga et al. [14••] reported an increase in detection accuracy of 11% against the best-performing image analysis ITDCD approach, "edge detection and region growing", using satellite RGB images. Pulido et al. [47] reported a 16% improvement in F1-score when comparing the best-performing CNN model with its counterpart from traditional methods, local maximum filtering, using multispectral images, and elevation models. Zheng et al. [48•] compared a customised CNN model and five other CNN models designed for object detection against random forest and support vector machine for tree detection using RGB images. The result showed the CNN models outperformed two traditional approaches by at least 29.76% and 23.02% in F1-score in two study sites.

Data Type

RGB imagery was the most frequently used data type in the reviewed studies (26 out of 35 studies). One reason is that many CNNs were initially designed for 3-channel inputs and pre-trained with large RGB datasets for computer vision tasks [31]. Although the domains of pre-training data are usually different from ITDCD, the knowledge learned through the process still gives the model the ability to extract general object features from RGB images and can effectively reduce training costs [24••]. As a result, most reviewed studies used RGB images directly without modifications to the data or the model structure. A few studies adopted pre-processing methods such as cloud removal [49], pan-sharpening [14••, 50], or band swapping [51]. Those methods are designed for certain image types and forest conditions and therefore may not necessarily be required for all RGB datasets.

When using multispectral images or lidar data, adjustments are required to handle higher dimension inputs than standard RGB images. These adjustments can be categorised into data dimension reduction and model structure modifications. The former approach simplifies high-dimensional data to accommodate the 3-channel input expected by pre-trained CNN models, while the latter approach modifies the architecture of a CNN model to allow direct input of higher-dimensional data.

In the reviewed studies, dimensional reductions are mostly found when processing multispectral images. A common process is to create a new band, named normalised difference vegetation index (NDVI), from near-infrared (NIR) and red bands. NDVI will then be duplicated three times to form a 3-channel image. In traditional ITDCD applications, NDVI has been widely used for forest area extractions [52–54]. In CNN-based applications, however, the benefits of using NDVI are varied when adopting different training and inference strategies. Mo et al. [55] found NDVI images produced a lower accuracy than RGB images when training CNN models separately on these two datasets. On the other hand, Safonova et al. [56] found that a CNN model trained with NDVI, RGB, and the green normalised difference vegetation index (GNDVI) produced greater or equivalent accuracies than using each of those data types alone. In another study conducted by Ampatzidis and Partel [57], NDVI was not used in crown detection but in a post-processing step to segment crowns within bounding boxes. Apart from deriving vegetation indices, other approaches have been used to reduce the dimensionality of multispectral images into 3-channel inputs. Xi et al. [43] compared several methods for band selection and band merging. They found the standard false-colour composite integrating red, NIR, and green bands produced the most accurate result. Pulido et al. [47] derived a new index band from multispectral imagery, named the digital elevated vegetation model (DEVM) which merges 2D spectral bands with 3D information from structure-from-motion (SfM) techniques. The DEVM band was then converted to a 3-channel image (DEVM × 3) for training and inference. A similar process was also applied to lidar data by Windrim, Bryson [58], who converted 3D lidar point clouds to 2D surfaces and then stacked them into 3-channel inputs for ITDCD.

In contrast to reducing data dimensionality, modifying the model structure can be used to enable direct use of higher-dimensional data and hence retain more of the original information for the model to learn from. This can be achieved by extending the first layer of the CNN model to accommodate additional input channels. The initial weights of those extended channels are usually copied from pre-trained RGB channels to reduce training efforts. Using this approach, Park et al. [41] created several CNN models that used different image products derived from multispectral data. Their result showed that adding multispectral information generally produced greater accuracies than using RGB only when detecting trees affected by pine wilt disease (with 9.9 to 15.67% increases in mean average precision) but produced similar results in dead tree detection. Hao et al. [59] adopted a similar approach to include both 2D spectral bands and 3D information derived by SfM from multispectral imagery. The resulting NDVI + CHM model produced the highest accuracy, with 7.02 to 12.57% increases in F1-score.

High-dimensional inputs also result when using data from multiple sources. In this case, dimensionality reduction and model architecture modification may suffer from two major issues: data misalignment and redundant learning. The former issue usually exists when two data sources are collected from different platforms or sensors, for example, horizontal misalignments between lidar and aerial imagery. The latter issue is caused by irrelevant or noisy information from additional bands, which can disturb the learning process and eventually reduce the accuracy [60]. In contrast, Pleşoianu et al. [42] proposed an ensemble model that uses a voting strategy to handle multisource data (e.g. lidar and RGB). Their model consists of several independent CNN models trained with raster products derived from both data sources. The final predictions are made by calculating the "votes" from those independent models, which avoids directly processing those bands through a single model. In comparison to the RGB models, they found the ensemble models provided respective improvements in detection accuracy of 3.33% and 7.09% in plantation and urban forest study sites.

In contrast to image analysis techniques, which are often more successful using lidar data, CNN models have been primarily developed for RGB imagery. However, multispectral, lidar data fusion has also shown potential utility in ITDCD applications. When using these high-dimensional data sources, adopting an appropriate approach is important. Data dimensionality reduction is fast and simple to apply but requires a careful selection of the methods for band preprocessing, training, and inference. The newly created data may also yield poor results when using an RGB pre-trained model and a small number of training samples [55]. In comparison, model modification has less information loss from a data perspective but can be technically challenging [61]. Despite requiring greater training effort, modified models generally benefit from using multispectral data, especially in areas where RGB models produced lower accuracies (e.g. diseased tree detection). This review identified only a single study [42] that used data from multiple data sources, and future research should use the approaches identified in this section.

Platforms and Resolutions

UAVs were the most common platform for data collection in this review (24 out of 35, with one study using both UAV and aerial images captured from manned aircraft). Compared with other platforms, UAVs usually fly at low altitudes, typically below 500 m above ground level, which leads to a smaller capture extent per flight but very high-resolution (VHR) images with ground sample distance (GSD) less than 5 cm. In the reviewed studies, UAV images have a range of GSD between 0.35 and 16 cm and an average GSD of 4.69 cm. For other aerial platforms including higher-altitude aeroplanes and helicopters, the GSDs range between 10 and 30 cm, with an average GSD of 15 cm. Two studies that used images from satellite platforms have GSDs of 50 cm.

Since CNN models rely on features extracted from images for different tasks, low-resolution images may not contain sufficient detail of objects such as object boundaries and textures. Therefore, it is unsurprising that VHR images were the most common choice in the reviewed studies. Fromm et al. [62] compared three CNNs on images at four different GSDs (0.3 cm, 1.5 cm, 2.7 cm, and 6.3 cm) for conifer seedling detection and found lower-resolution images decrease mean average precision by between 13 and 15%. When using 0.3 cm images, larger seedlings also showed higher detection accuracies than smaller ones, which benefitted from more visible and distinctive features. Another study conducted by Ocer et al. [63] found a decrease in precision by 17% and F1-score by 9% for tree detections in an urban area after downscaling the testing image resolution from 4 to 6.5 cm. Zheng et al. [48•] detected oil palm trees from two study sites using imagery with GSDs of 10 cm and 8 cm, respectively. The results showed that the F1-score was 11.13% lower in the site with 10 cm GSD imagery.

On the other hand, the advantages of high-resolution imagery may not be obvious when tree features are more distinguishable. Safonova et al. [56] found little impact of resolution changes on the detection of oil palm trees in a sparsely planted area. After decreasing GSDs from 3 to 13 cm, the F1-score of crown delineation only dropped slightly, by 0.42% and 0.86% for the models with and without data augmentation, respectively. The unique shape of the trees and their contrast with background in this environment may explain this lack of sensitivity to image resolution.

In addition to VHR images from UAVs, several studies also proved the feasibility of using lower-resolution aerial imagery (10–30 cm) for ITDCD [30••, 38, 39, 64]. However, it is difficult to summarise the minimum GSD by forest type required to achieve a certain level of ITDCD accuracy because there is a lack of statistical description of crown characteristics (crown size distributions, colour histogram, shape, etc.) in most of the reviewed studies. Future studies could be conducted to explore the relationship between tree characteristics and imagery resolution to identify potential interactions.

In addition to image resolution, input image size is another factor that can affect the accuracy of CNN models for ITDCD. Due to limitations on computing memory, most studies split large remote sensing images, usually over 5000 × 5000 pixels, into smaller input images, usually between 64 × 64 pixels and 1280 × 1280 pixels, for training and inference. The sizes used reflect computer memory capacity as this provides CNN models with a wider field of view. From an inference perspective, an obvious advantage of larger image tiles is allowing wider overlap between images when clipping image tiles, which increase the chance of capturing the complete crown. In addition, a larger image size may also be required when applying a trained ITDCD model on coarser resolution images as it will maintain the ratio of tree objects in the images and offset the impacts of resolution reduction [64].

On the other hand, large tile size can be detrimental when target objects are relatively small [65]. In the case of forest environments where there can be many small crowns, a medium-sized tile may offer better performance in terms of detection accuracy and computational efficiency. For example, Culman et al. [39] tested three image sizes between 64 × 64 pixels and 256 × 256 pixels for palm tree detection and found that 128 × 128 pixels produced the best result.

Forest Environments

The reviewed studies were conducted on plantation forests (15 articles), natural forests (9 articles), urban forests (6 articles), and mixed forests (5 articles). In contrasting the studies on these differing forest types, the performance of CNN models was generally affected by three factors: stand density, crown characteristics, and background conditions. Under dense forest conditions, closely adjacent or overlapping crowns made edge detection on individual trees more challenging, especially between those trees with similar canopy appearances. Dense stands also resulted in more overlapped and shaded crowns, which many CNN models struggled with [14••, 27, 41, 49, 51, 55, 66, 67].

With respect to crown characteristics, variations of size, shape, and colour can reduce model performance. This is usually caused by two factors. First, training samples failed to cover sufficient crown variation. For example, in urban areas, where species richness can reach hundreds of species [68], it may be difficult to incorporate sufficient training samples to describe the variation in those species' crowns [39, 69•]. Second, remote sensing imagery failed to provide sufficient information to distinguish variations in crown characteristics. Low spatial or spectral resolution imagery may result in a CNN model's inability to identify small crowns or separate large variability within crown types [48•, 62–64].

The background, or surroundings, for trees is another factor that can influence the performance of CNN models for ITDCD. Some complex forest environments may contain tree-like objects in the background that are detected as false positives. These could include shrubs [29], weeds [57], other low-lying vegetation [30••, 48•, 63], or even non-vegetated features that share similar visual characteristics with trees, like streetlamps [70]. Heterogeneous backgrounds also

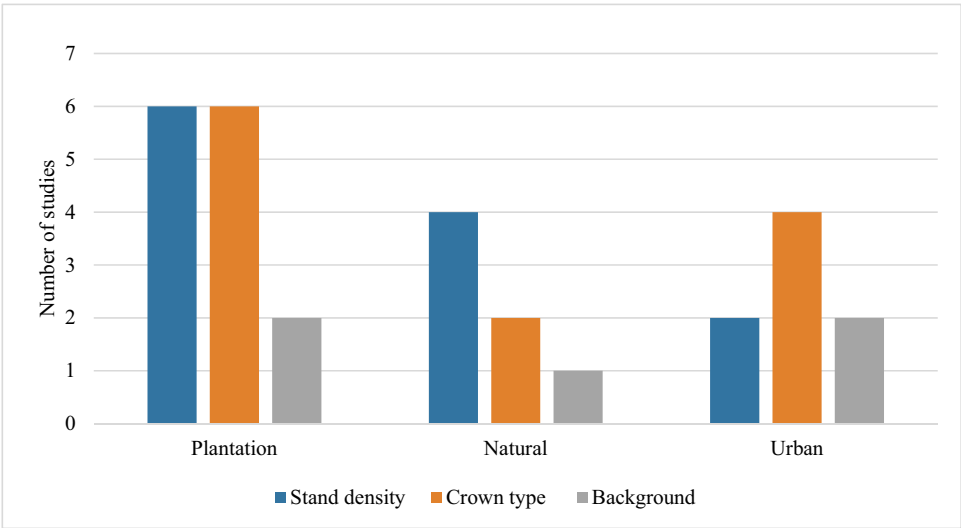

Fig. 1 Error sources interpreted from the reviewed studies and grouped into three main categories

increase variation in tree appearance, which results in more samples being required to effectively train the CNN [39].

Among the three major forest types (plantation, natural, and urban), 23 studies provided analysis of major error sources (Fig. 1). Plantation forests are managed sites and therefore tend to report fewer errors from background noise in contrast to crown characteristics and stand density. Crown characteristic errors in plantation forests are mostly caused by under-sampled variation, including smaller crowns and different growing conditions [28, 48•, 62], unique species [58], and unseen tree forms [50, 71]. Stand density errors are also commonly reported, including overlapping crowns [27, 28, 55, 59, 66] and errors from shadows [48•].

Natural forests usually consist of closed canopy areas with dense stands and complex vertical structures, which result in similar stand density problems to plantation forests. The reported errors include overlapping crowns [14••, 51, 67] and shadows [49], as well as under-sampled crown variation [30••, 71].

Urban forests contain a considerable variety of tree forms, species, sizes, and groupings. Also, they have a greater variety of background features including many vegetation types and man-made features. Since most man-made features are visually distinguishable from trees, those features were not found to have major impacts on the accuracy of ITDCD with CNN. Instead, vegetation features such as shrubs [63] and grass [70] were found to cause errors in some studies. For urban forests, the outliers in crown size are more commonly reported issues. Zamboni et al. [69•] found large trees are likely to be split into separate crowns or only partially detected. Smaller trees were reported by Ocer et al. [63] and Xia et al. [70] as error sources, while Xi et al. [43] pointed out that both large and small trees caused detection errors.

Unlike traditional ITDCD approaches, which usually treat tree detection and classification as separate tasks [16, 18], a CNN model can integrate both steps by extracting bounding boxes or crown masks for trees. From this point of view, the ITDCD models can be further grouped into three categories: "single class", which detects crowns of a single species; "multiple classes", which detects crowns for multiple species; and "general class", which detects all trees in a study area without further classification. Extraction of general tree objects or tree crowns with multiple classes is more difficult than the detection of crowns of a single tree class. The former requires a model to be generalisable for detecting varied characteristics of crowns, while the latter requires the model to be able to distinguish certain classes of tree crowns from canopy areas. Both tasks require numerous training samples in order to train the model for various situations [30••, 66, 72].

Extracted classes tend to be different across forest types (Table 2). Most plantation forests are managed monocultures, and therefore single-class detections are more commonly applied in these environments (8 out of 15). In comparison, deriving a single class of trees from other forest types is more challenging due to interference from surrounding trees and backgrounds. As a result, most natural forest or urban forest studies either choose a tree class that has a unique appearance (e.g. palm trees or diseased trees) or are undertaken in a small study area (less than 15 ha) with a homogenous environment, relative to the broader area. For general class detections, a need for numerous high-quality training samples was reported by studies from both urban [40, 63, 69•] and natural forests [14••, 30••]. Although the monoculture setting in plantation forests should make general detection less challenging, two studies [28, 58] concluded that including more variation in training samples would be crucial for further improvements in accuracy in these environments. Studies using multiclass detection were rare. In plantation, natural and urban forests, multiclass detection was only conducted to differentiate classes within a single species, for example, palm trees with different health conditions. The only study achieving species classification was by Pleşoianu et al. [42], who used CNN models for ITDCD of plum, apricot, and walnut trees from a horticultural plantation area and coniferous and deciduous trees from a natural forest area.

In summary, stand density, crown characteristics, and backgrounds are three factors that affect the performance of CNN models under different forest types. Some proposed solutions such as adding more training samples, using higher-resolution images or data fusion may not be necessary for one forest type but crucial for another. Identified research gaps include investigating multispecies ITDCD, single species ITDCD (other than diseased trees) within complex forest environments, and the impact of the number of training samples and their quality on general class detection tasks.

CNN Models

As mentioned in the introduction, CNN-based crown detection and delineation refer, respectively, to object detection

Table 2 Number of studies categorised by forest types and crown detection methods

	Plantation	Natural	Urban	Mixed
Single	9 [27, 29, 50, 55–57, 62, 71, 73]	3 [38, 51, 74]	3 [43, 70, 75]	1 [39]
Multiple	2 [48•, 66]	3 [41, 49, 72]	1 [40]	1 [42]
General	4 [28, 47, 58, 59]	3 [14••, 30••, 67]	2 [63, 69•]	3 [64, 76, 77]

and instance segmentation in computer vision. In general, CNN models for those tasks consist of two main components, a "backbone" architecture that extracts features from input images and a "detector" that generates bounding boxes or per-instance masks for the objects (crowns). The following sections review both components of CNN models and their applications in ITDCD studies.

Backbones

The backbone or feature extractor network is the key component of a CNN model. After the successful applications of early CNN networks such as LeNet and AlexNet on image classification, later published networks started to be used as feature extractors (backbones) for object detection and instance segmentation. In these models, deep convolutional neural networks first extract object features at multiple levels from input images and then feed the features into the detector where bounding boxes or masks of the objects are predicted. As a result, the backbone's ability to represent and extract features directly affects the performance of the detector and therefore the accuracy of the model. Table 3 summarises the highlights of various backbones.

ResNet was the most commonly used default backbone in the reviewed studies. Compared with previous backbones, ResNet improves feature extraction significantly by introducing a deeper but more efficient network design. ResNet has gained widespread popularity in various applications [84, 85]. However, the benefit of using deeper ResNet backbones for ITDCD tasks remains unclear. For example, Fromm et al. [62] reported an improvement of 15% in mean average precision when using ResNet-101 rather than ResNet-50 for conifer seedling detection, while Weinstein et al. [30••] found that a deeper ResNet did not provide significant improvements when used for general tree detection. Despite its popularity, ResNet has been reported to produce lower accuracies when compared with other backbones under the same CNN framework. For example, ResNet yielded lower accuracies than Inception v2 when using the same Faster-RCNN detector for crown detection from a vegetation index image [47]. The performance of ResNet and Inception v2 backbones was similar for leaf-on detection of almond trees, but ResNet resulted in lower accuracy when applied to pine trees. In addition to ResNet, some other backbones have also been applied to ITDCD tasks, including EfficientNet [40, 49] and DarkNet [57, 74]. However, those backbones were designed to fit specific CNN detectors, which are normally not compatible with ResNet backbones. Therefore,

Table 3 A summary of highlights for the various backbones in this review

Backbone	Reference	Highlights
VGG	Simonyan, Zisserman [78]	• Deeper network for more complex feature extraction (16 ~ 19 layers) • Smaller filter and stride in convolutional layers with fewer parameters
Inception (GoogleLeNet)	Szegedy et al. [79]	• A stack of local networks named the inception module • Each module has parallel convolutional filters and bottleneck filters, which reduce input detections and improve computational efficiency • Deeper network with 12 times fewer parameter than AlexNet • Later versions, Inception v2 and Inception v3, have further improvements and reach higher accuracies
DarkNet	Redmon et al. [80]	• Specifically designed for You Only Look Once (YOLO) detector • Combines some features from Inception and VGG • Later versions include DarkNet-19 and DarkNet-53, which have more efficient convolution layers and residual blocks for higher accuracy
ResNet	He et al. [81]	• Popular backbone that works well for a range of tasks including object detection and instance segmentation • Significant increase in layer depth (50–152 layers) • Residual block and skip connection designs allows efficient optimisation even with very deep networks
EfficientNet	Tan, Le [82]	• Scalable network design allows dynamic adjustments to structure using a neural architecture search technique • More flexible and efficient when extracting features under different scenarios • Some evidence for advantages in ITDCD tasks when compared with ResNet and Inception [40, 49]
FPN (feature pyramid network)	Lin et al. [83]	• A top-down, image pyramid network that can be used with other backbones for integrating multilevel features • Combine low-level feature maps that are rich in location information with high-level features that contain greater semantic information • Predictions are made from the integrated feature maps at each up-sampling level

comparisons between those models and ResNet were made only at the model level.

Pre-Training

Training a CNN is computationally expensive and requires a large number of labelled samples [24••, 86]. These samples are usually created by manual annotation (labelling) of tree crowns or canopy cover using the remote sensing data as a reference (e.g. RGB, multispectral, lidar). In the training process, the backbone or entire CNN model needs to adjust the weights on every neuron across the network for predictions of input samples [85]. To achieve accurate detections, the training samples must cover sufficient variation to allow the model to extract general features at different levels of the network (avoid underfitting) but without memorising the data (avoid overfitting) [87, 88].

For most object detection tasks, including ITDCD, acquiring or training sufficient ground samples to train a model is not feasible or required. Instead, a common approach is to use backbones or models that have been pre-trained on very large, annotated datasets—often from a different domain. Those datasets usually contain a range of objects and can give a model the ability to learn to extract general low-level features common across domains such as colour gradients, textures, and shapes. When training for specific object detection, a pre-trained model often only needs to be adjusted (fine-tuned) with samples from the target domain to enable the detection of the target objects. This can greatly reduce the computing resources and training data required to train large models [28].

In this review, most studies used models pre-trained on two large datasets: common objects in context (COCO) and ImageNet. The COCO dataset contains 91 types of objects with segmented instances [44], while ImageNet consists of tens of millions of annotated images that are categorised by a hierarchical structure. The most commonly used ImageNet data is a subset created for the ImageNet large-scale visual recognition challenge (ILSVRC) which contains over 1.4 million images across 1000 categories [89]. Here, we refer to that dataset as "ImageNet". A detailed comparison of the two datasets is described in Table 4.

Table 4 Summary of commonly used pre-train datasets

	COCO	ImageNet (ILSVRC)
Number of images	328,000	14,197,122
Instances	2.5 million	N/A
Image resolution	Varied	Varied
Categories	80 (2014 dataset) and 91(2015 dataset)	1,000
Training samples	82,783 + 165,482	1,281,167
Validation samples	40,504 + 81,208	50,000
Testing samples	40,775 + 81,434	100,000

Most studies reported the benefits of applying COCO or ImageNet pre-trained models for ITDCD, while some studies further assessed the model with additional samples on different detectors. Chadwick et al. [28] compared Mask-RCNN models (see Detectors section) pre-trained with two versions of the COCO dataset for crown delineation. They found that the expanded version with images of balloons increased the F1-score by 12% when trained on a full model (detector head and backbone simultaneously). The study further argued that the model benefited from the similarities of geometry between tree and balloon objects. Fromm et al. [62] estimated the impacts of COCO pre-trained networks on three different detectors. When pre-training was applied, the deep backbone used in the two-stage detectors saw increases in accuracies, but the one-stage detector with a shallower backbone showed the opposite trend.

Moreover, Culman et al. [39] suggested that pre-trained models using general object images, such as COCO and ImageNet, should not be applied directly to remote sensing data. They found that image samples taken from a birds-eye view are essential for improving the model's accuracy. Instead of using a pre-trained model, Zheng et al. [48•] trained a modified CNN model with 363,877 manually collected palm trees. However, collecting sample sizes of that magnitude is not likely to be feasible in many instances. Thus, it is worth exploring the potential benefits of using large-scale remote sensing datasets as an alternative to COCO and ImageNet for model pre-training.

Detectors

CNN frameworks for object detection can generally be classified into two categories: two-stage detectors and one-stage detectors. Two-stage detectors consist of a region proposal module and an object detection/classification module. The regional proposal module first creates many possible bounding boxes using image analysis techniques or convolutional neural networks. Then, the second module is used to extract features from the proposed bounding boxes for classification and bounding box refinement [90]. Commonly known two-stage detectors include region-based convolutional neural network (RCNN) families and region-based fully convolutional networks (R-FCN). In contrast, one-stage detectors integrate the tasks of object classification and localisation of the bounding box or mask into a global problem and produce detections in one-stage [85, 90]. When training a CNN model, the detector will be optimised alone or together with the backbone using a loss function. Commonly used one-stage detectors include YOLO [80], RetinaNet [91], and single-shot detector (SSD) [92]. Table 5 summarises the highlights for detectors identified in this review.

Table 5 Summary of detectors used in reviewed studies and their highlights (for studies that used multiple models, the best performing model was included)

Detector type	Reference	Highlights	# of studies
One-stage			12
RetinaNet	Lin et al. [90]	• A novel loss function, named focal loss, that is less sensitive to redundant bounding boxes (such as noise bounding boxes from the background) • Uses FPN to enhance multilevel detections	6 [30●●, 39, 64, 66, 75, 76]
EfficientDet	Tan et al. [92]	• More efficient backbone (EfficientNet) • More efficient feature fusion network, named bi-directional FPN (Bi-FPN). The network uses two-way connection for cross-level feature integration and a skip mechanism • A significant reduction in parameters • Shows advantages in both processing speed and detection accuracy compared with other popular CNN models including Mask-RCNN, RetinaNet, and YOLO v3	2 [40, 49]
YOLO	V1:Redmon et al. [79] V2:Redmon, Farhadi [93] V3: Redmon, Farhadi [94] V4: Bochkovskiy et al. [95] PP-YOLO: Long et al. [96]	• A fast, one-stage detector that has developed into several versions • Produces fewer background errors compared with two-stage detectors as it can include more contextual information by processing the image as a whole • Very low computational cost and widely used for real-time object detections • Several enhanced versions were developed to overcome the shortcomings of predecessors	1 [57]
DetectNet	Tao et al. [97]	• Uses Inception network as backbone • Very fast single-stage detector • Uses a regular grid and 3-dimensional label data to locate all objects in an image and therefore more efficient training and inference	1 [47]
YOLACT	Bolya et al. [98]	• One-stage instance segmentation model • Faster than Mask-RCNN and can achieve real-time instance segmentation • Lower accuracy than Mask-RCNN in computer vision competition	1 [55]
BlendMask	Chen et al. [99]	• One-stage instance segmentation model • Outperforms Mask-RCNN in computer vision competitions on processing speed and accuracy • Improved merging process between low- and high-level features, which was not well-addressed in YOLACT	1 [43]
Two-stage			14
RCNN	Faster RCNN: Ren et al. [100] Mask RCNN: He et al. [101]	• Original RCNN was developed by Girshick et al. [102] and then further improvements were made in subsequent versions • Faster R-CNN and Mask-RCNN are very popular models for object detection and instance segmentation • Apart from ResNet, more up-to-date backbones have been used in RCNN structures	14 [14●●, 27–29, 38, 48●, 50, 51, 56, 58, 59, 62, 63, 70]
Others		Customised models (8) and anchor-free model (1)	8 [41, 42, 67, 71–74, 77] + 1 [69●]
Total			35

In general, two-stage detectors from the RCNN family were most popular, with eight studies using Faster-RCNN and RCNN for tree detection, and six studies using Mask-RCNN for individual crown segmentation. The one-stage detector, RetinaNet, was used in six studies and was the second most popular model. When comparing two types of detectors, two-stage detectors are usually slower than one-stage detectors, largely because of the additional regional proposal network. The advantages of two-stage detectors were obvious when RCNN models were first introduced between 2014 and 2017, as the more complex network structure generally resulted in higher detection

accuracy. However, more recently published research shows one-stage models have started to overtake two-stage models in some computer vision tasks in recent years [93, 94].

This trend was also observed in ITDCD applications. For example, Dos Santos et al. [75] found that one-stage detectors, RetinaNet and YOLOv3, showed advantages over Faster-RCNN when detecting baru trees (*Dipteryx alata*) from very high-resolution images. The one-stage detectors produced higher average precision than Faster-RCNN by 10.16% and 6.76% with, respectively, 6.3 and 2.5 times faster processing speeds. Ammar et al. [40] compared EfficientDet with Faster-RCNN, YOLOv3, and YOLOv4. EfficientDet produced the highest average precision with increases ranging from 5.14 to 9.57% on palm tree detection and 3.38 to 34.03% on general tree detections. Huang et al. [49] undertook a similar comparison of these four models for the detection of pine trees infected with wilt disease. EfficientDet also produced the best results with increases in average precision from 1.34 to 3.07%. For individual crown delineation, some novel one-stage detectors have also been found. Mo et al. [55] used a lightweight one-stage detector, YOLACT for lychee (*Litchi chinensis*) tree crown delineation, and produced a mask AP of 95.44%. Another one-stage detector, BlendMask was compared with Mask-RCNN for delineation of ginkgo tree (*Ginkgo biloba* L.) crowns and showed an increase of 6.6% for AP.

In a study by Zamboni et al. [69•], an anchor-free CNN was used for ITDCD in an urban forest area. The study reported that the anchor-free detector outperformed both one-stage and two-stage models and produced the highest average precision (with 1.7–9% improvements). Compared with one- and two-stage detectors, which start object detection from a large number of pre-proposed regions (or anchor boxes), anchor-free models directly predict the key points (i.e. corner points of bounding box) or centres of the object. This simplified structure not only improves computational efficiency but also makes the model more adaptable to different object scales as there are no limitations from the proposed regions [95]. These advantages may explain the improvements as the study area contains crowns at varied scales and the size of training samples is relatively small.

It should also be noted that the performances of a model may not persist across different forest conditions. Emin et al. [51] compared Faster-RCNN, with three one-stage detectors, YOLOv3, YOLOv4, and SDD, for spruce (*Picea schrenkiana*) crown detection in areas with different tree densities. Despite a longer training time, Faster-RCNN outperformed the one-stage detectors, exhibiting a significant increase in overall accuracy from 13.94 to 39.27% across test areas, with greater improvements occurring in images with lower crown densities. Faster-RCNN showed comparable performance with a one-stage detector, DetectNet, on individual tree detections from a leaf-on almond tree (Prunus dulcis) site, with less than a 5% difference in F1-score [47]. This gap became much larger when testing both detectors on another site with pine trees (*Pinus greggii*) that had a higher proportion of overlapping crowns, with reductions in F1-score between 22 and 42%. There are two possible reasons for this difference. Firstly, those models used a unified parameter setting, which may suit one environment but degrade the performance in another environment. Secondly, one model may have a lower learning efficiency compared with the others using the same training data.

Customised Structures

Instead of using models developed for computer vision tasks, several studies further modified model structure to fit ITDCD tasks. Most of those changes focused on resolving specific issues including reducing shadow impacts, enhancing small tree detection, and improving computational efficiency. Zheng et al. [48•] proposed a customised method based on the Faster-RCNN structure for the detection of five oil palm statuses: healthy, dead, mismanaged, smallish, and yellowish. The changes included adding a refined pyramid to enhance small oil palm detection, adjusting anchor size and aspect ratio, and changing the loss function. These changes increased accuracies by a significant margin between 8.14 and 21.32% compared with the original Faster-RCNN and five other CNN structures. Zhou et al. [71] developed a new model based on VGG and single-shot detector (SSD) by adding more efficient convolution layers and a feature enhancement pyramid, which produced better detections for small diseased pine trees (*Pinus tabuliformis*). Ye et al. [67] added a generative network on Fast-RCNN to create masks on shadow areas and enhance tree detections. The modified model showed an improvement of between 2.2 and 5.7% compared with the original Fast-RCNN model when detecting trees from forests with three stand densities. Li et al. [74] modified YOLOv4; the resulting lightweight version of it was able to run on an edge computing platform for initial image selection. The filtered images were then sent to a ground station for more accurate detection.

In addition to object detection and instance segmentation models, Tong et al. [73] used a semantic segmentation CNN model, U-Net, to first extract canopy areas and then apply a rule-based function for individual crown segmentation. However, the method was developed for a plantation area that contained very few overlapping crowns and is not applicable for complex forest environments like those in urban areas. A similar approach was also used by Ferreira et al. [72], who combined a CNN semantic segmentation model with a morphological operation for the delineation of three palm species *Attalea butyracea*, *Euterpe precatoria*, and *Iriartea deltoidea*. Due to the unique shape of palms, this

hybrid model also has limitations for tree detection in other forest environments.

Training Strategies and Techniques

Training Samples

Most studies collect training samples via hand annotation, where individual crowns are manually labelled using bounding boxes or crown masks. The number of samples varied greatly between reviewed studies, ranging from 110 [75] to 363,877 [48•]. This makes it difficult to identify a minimum sample number for general use in ITDCD tasks. Since CNN models are generally data hungry [96], increasing the number of samples can bring benefits to the training and accuracy. However, quality and sample balance are also important perspectives that should be considered to avoid overfitting and increase generalisability [97].

In addition to hand-annotated samples, several studies developed semi- or fully-automated methods to generate large numbers of synthetic training data. These methods have significantly increased the efficiency of model training and improved the generalisability of the models for tree object detection. Weinstein et al. [30••] proposed an unsupervised method to automatically segment tree crown shapes from lidar point clouds. The RGB samples clipped by those shapes were then used to train a CNN model. Compared with hand-annotated data, their method produced a large number of samples (434,551 crowns) within a very short time. This method was also used in his later studies [64, 76] to create training datasets containing over 9 million and 30 million trees, respectively. Although the model was further fine-tuned with small numbers of hand-annotated data (2,848 crowns), the initial weights were learned in a more efficient way. Other studies also created synthetic training images from a small number of hand-annotated data. Chiang et al. [38] created 5,000 tree samples using 300 hand-annotated trees and 63 background images. Braga et al. [14••] created 19,656 synthetic images that contained 4 to 150 trees using 1506 hand-annotated trees. Pulido et al. [47] developed a morphological algorithm to simulate the shapes of tree crowns in a digital elevation canopy model. The study produced 12,500 synthetic images that contained 2 to 7 trees without creating any hand-annotated data. Some of those methods applied a copy-paste strategy when creating the samples. This strategy can be considered as a data augmentation process using existing samples and is discussed in more detail in the following section.

Data Augmentation

Data augmentation is a method for expanding samples from a limited number of existing data [98]. The method is effective in avoiding overfitting problems by adding more variance to the sample data, which is usually achieved by performing operations such as geometry and colour space transformations, random erasing, and kernel filtering on the input images before feeding them into a model for training. Those operations can be divided into position augmentation and colour augmentation.

Nearly half of the reviewed studies (16 of 35) applied these data augmentation operations (Fig. 2). Position augmentation that simulates variances of tree crown shape, scale, and orientation was used in 15 studies. In comparison, colour augmentation was only found in six studies, all in combination with position augmentation. A possible reason is that colour augmentation simulates variance in the

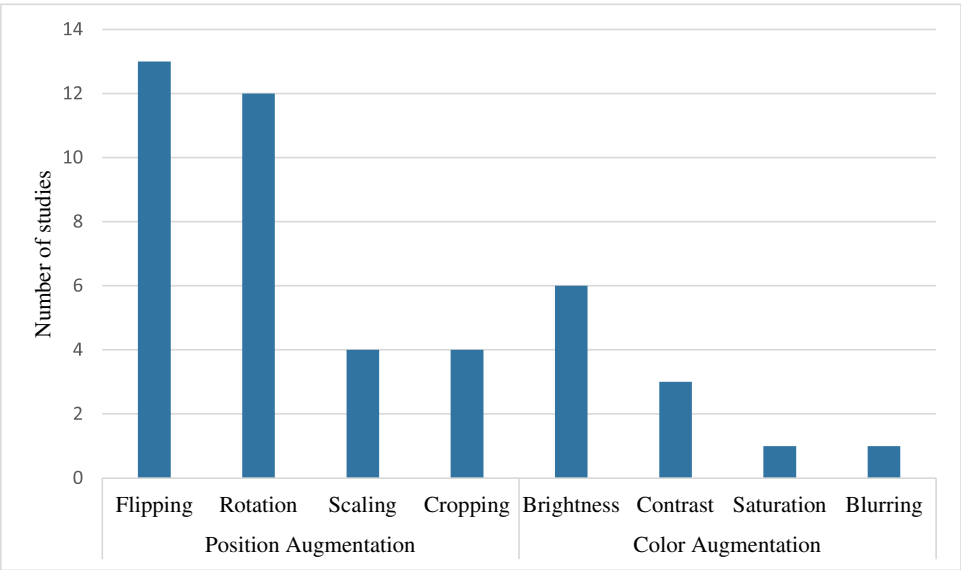

Fig. 2 Position and colour augmentation techniques used in reviewed studies

images such as illumination conditions and spatial distortions, which may not be necessary for interpreting images with consistent qualities. For position augmentation, flipping and rotation were the two most common techniques used in 13 and 12 studies respectively, followed by scaling and cropping, each used in four studies. For colour augmentation, brightness adjustment was most often applied ($n=6$), followed by adjusting contrast ($n=3$), saturation ($n=1$), and image blurring ($n=1$).

The impact of data augmentation on ITDCD accuracy is inconclusive in the literature reviewed. Gomez Selvaraj et al. [66] claimed that basic data augmentation provided a positive impact on the model. Pleşoianu et al. [42] argued that applying data augmentation helped to reduce the risk of overfitting. Braga et al. [14••] tested several combinations and found the augmentation composed of flipping, rotation, and saturation produced the highest accuracy using Mask-RCNN. On the other hand, Weinstein et al. [76] found that random flips and translations did not improve the model accuracy when trained with images generated from hand annotations and unsupervised detections. Fromm et al. [62] found increases in accuracy on SSD and R-FCN detectors when applying flipping, cropping, and rotation, but no significant improvements were observed for the Faster-RCNN detector. The study pointed out that the contexts covered by training images would be important for identifying new images. Safonova et al. [56] found that simple augmentation operations decreased the Mask-RCNN accuracy since it produced new samples that were too similar to other objects. In summary, the impacts of image data augmentation are dependent on CNN models, technique combinations, forest conditions, and data types.

In addition to augmenting the existing samples, three studies used techniques to generate synthetic tree samples. These methods are rule-based and designed by researchers' subjective understanding of tree features. Compared with hand-annotated data, these manipulation methods have the potential to artificially create more variance in the sample data and therefore improve the generalisability of the CNN model. A detailed comparison of the methods is shown in Table 6. Chiang et al. [38] created synthetic training images using samples of dead tree crowns from RGB imagery. The hand-annotated crown samples were first processed by image augmentations and then randomly placed on the background of images that did not contain dead trees, a so-called "copy-paste" technique. The method reported a relatively low accuracy with a 34% F1-score. Despite few false positives (commission errors), many omission errors were observed. Braga et al. [14••] adopted a similar copy-paste strategy with RGB imagery. However, instead of placing trees randomly, they added several user-defined parameters to control the number and density of trees in the synthetic images. The study achieved good overall accuracy, with an F1-score of 86%, but still reported a major error source from unseen crown variation. A study conducted by Pulido et al. [47] used a different approach and dataset. They proposed an algorithm to simulate the conical shapes of tree crowns and created synthetic training images from a digital elevated vegetation model (DEVM), a representation that combines multispectral images, digital surface models, and digital terrain models. The highest F1-score achieved by the CNN model was 94%. However, since the approach requires multispectral and photogrammetric images to create the DEVM, applying the method in new areas would require the availability of the same data types and the derived DEVM layer.

Table 6 Details of studies using copy-paste synthetic sample methods

Study	Key techniques used in the method	Objective	Hand -annotated samples	Synthetic images	Accuracy (F1-score)
Chiang et al. [38]	• Image augmentation: reshape, rotate, brightness adjustments • No control on tree numbers and density	Dead tree, natural forest	300	5000	34%
Braga et al. [14••]	• No image augmentation • User defined parameters to decide number of trees and crown density	General tree detection, natural forest	1506	19,656	86%
Pulido et al. [47]	• Morphological simulation algorithm to create conical crowns in the digital elevated vegetation model. The algorithm used a set of randomly defined rules to simulate the morphology of tree crowns	General tree detection, plantation forest	None	12,500	94%

Conclusion and Future Perspectives

In this paper, CNN-based ITDCD studies were reviewed through six themes. Each theme provides a summary of major trends and discussions on the factors that shaped those trends. Despite the generally impressive results achieved by CNN models, some research gaps are identified below.

Data Fusion

The potential for multimodal data fusion has not been well-explored. Combining RGB images with additional spectral or structural information could be useful in resolving errors caused by overlapping tree crowns, shadows, and background noise. The current methods for data fusion typically fall into two categories: low-level fusion (data dimension reduction, input layer modification) and high-level fusion (ensemble model). Both have their limitations; low-level fusion struggles with poorly aligned data, while high-level fusion processes each data source separately and impedes the optimisation of ITDCD as a global task. In comparison, middle-level fusion, or feature fusion, could offer more advantages by combining feature maps extracted at mid-to-late stages from the CNN backbones. As the features extracted at mid-to-late stages from a backbone are less local, the model is less likely to be affected by misalignment issues [99]. The feature fusion also means the entire model can be optimised and jointly learn to maximise the use of information from multiple data sources [100].

Improving Data Efficiency

Sample scarcity is a major issue that restricts both further improvements in accuracy and wider application of CNN models to more complex ITDCD scenarios (e.g. cross-site detection on different tree species). Apart from collecting more data, several approaches could be investigated to further improve data availability and efficiency.

Applying customised sampling strategies

Most reviewed studies selected samples in a random way and did not consider forest conditions. This random approach is suitable for plantation areas with homogeneous canopy conditions but may not be effective in complex forests that contain more crown variation. Instead, sample selection should be based on unique characteristics of the target context (forest setting) of interest. For example, increasing the number of shaded trees may help the model to recognise crowns under low illumination conditions, which are very common in natural and urban forests.

Dealing with an Imbalanced Dataset for Multiclass ITDCD

Balancing sample collection is an important factor to consider for multiclass ITDCD. Imbalanced datasets could result from the uneven distribution of tree classes in the real world, for example, endangered species versus others. Since the model has seen more instances of oversampled classes, it may become overfitted [97]. In this review, this issue was only discussed in one study [48•], which modified the loss functions to make the model learn equally across an imbalanced number of samples. Future studies could explore the impacts of imbalanced samples and test other methods at the data level (e.g. batch balancing, synthetic approaches) or at the algorithm level [101] to improve multiclass ITDCD.

Expand Sample Datasets Using Augmentation and Synthetic Techniques

Expanding the training dataset from hand-annotated samples is another way to improve data efficiency. The current studies did this through two approaches, augmentation, and the copy-paste synthetic sample approach. For image augmentation, further investigation is needed to analyse the effectiveness of different combinations of techniques. For the copy-paste approach, more controls could be applied to better simulate realistic forest conditions and scenes. An example of this could be adding shadow effects to create the illusion of overlapping crowns within groups of trees. Additionally, it is worth exploring the use of deep learning-based augmentation. This type of approach uses deep learning models to learn the distribution of features from real images and then uses that knowledge to generate new images. It has been demonstrated in a wide range of tasks as an effective way of improving data variability and reducing overfitting [87, 98, 102–105].

Model Selection

In the past few years, many CNN models have been developed for computer vision tasks. While many CNN models have improved both accuracy and computational efficiency, these improvements were developed using large datasets like ImageNet or COCO. In contrast, for tasks related to ITDCD, the training sample size is usually much smaller. Therefore, model selection should also consider learning efficiency, that is, which model can produce the best result with limited tree samples and computing resources. This is important because the sparsity of high-quality ITDCD training data might require a different strategy than simply selecting the state-of-art model from the latest literature or benchmarks.

From this point of view, choosing a deeper or more complex backbone may not be the best option in some cases. A lighter, shallower network is easier to fine-tune and hence

more likely to benefit from training with a small number of samples [106]. Moreover, a lighter network requires fewer computing resources and can leverage larger image tiles and epochs. An epoch refers to a single pass in which the CNN model is trained on the full set of available training images. By using more images per epoch, the model is able to derive more complex representations and prevent overfitting [107]. In addition to those factors, future research could also explore how feature extraction varies between model backbones. An intuitive method is to compare feature maps extracted from different stages of networks. When trained with the same number of samples, some backbones may produce better representations of tree objects and eventually achieve greater accuracy.

For detectors, some recently published models (e.g. BlendMask, EfficientDet, anchor-free detectors) have shown their potential utility for ITDCD tasks when compared with popular RCNN and RetinaNet models. Although many comparisons have been made, there is no single detector that outperforms the others across all conditions. These results indicate a wider evaluation of CNN models is needed to better understand how model characteristics (e.g. good at large-scale object detection) affect performance under different forest conditions.

Pre-Training

Most reviewed studies used backbones pre-trained with computer vision datasets (COCO or ImageNet). However, CNN models may produce better results when initialised with remote sensing datasets or crown-shape-like datasets (e.g. balloons and plants). These datasets will allow a backbone to learn more about the extraction of tree features from an aerial view rather than general objects taken from oblique view angles. Some sampling approaches, such as crown synthetic models and an unsupervised method proposed by Weinstein et al. [30••], could also provide a large number of samples for initialisation with low labour efforts. Future studies should explore the impacts of those initialisation datasets.

Model Training

Although comparisons between model structures were conducted in previous studies, little attention has been given to exploring the optimal parameter setting for each individual model. Unified parameter settings are usually applied when comparing models whose structures are fundamentally different. Selections of the unified parameter settings were also based on a random search approach, in which groups of parameters were subjectively chosen by the user and then compared using a small number of training samples. This comparison strategy may be able to answer which model produces the best result under a certain parameter setting. However, another question to investigate is whether one model can outperform another given the same level of effort put into parameter optimisation. Parameter search methods such as Bayesian optimisation, hyperband optimisation, or grid search could be applied to quantify parameter search efforts and assess a model's performance [108•].

Assessment of Crown Detection and Delineation

The extent of tree crowns is an important parameter in forest management and is useful for estimations of tree growth efficiency and stand competition [109]. The bounding boxes extracted by CNN models would be a useful resource to measure crown spread if those boxes accurately aligned with the crown shape. However, most crown detection studies did not assess bounding boxes from this perspective. Although some studies used IoU to evaluate the extent of predicted bounding boxes, the quality of ground truth samples in those studies varied. Annotations of individual crowns mostly focused on representing the locations of crowns rather than providing clear crown extents. For example, the bounding boxes created by Dos Santos et al. [75] only cover the major area of a crown since the primary purpose of their study was to detect the existence of individual trees rather than individual crown extents.

In summary, CNN models have shown considerable potential in ITDCD tasks, especially when using only RGB images, on which traditional ITDCD approaches generally fail to produce promising results. In some comparisons, such differences can range between 11 [14••] and nearly 30% [48•] in F1-score. Despite good results with only RGB data, it is still worth exploring the benefits of incorporating additional information (spectral, structural, and spatial) for ITDCD with CNN. A critical aspect of this exploration is selecting an appropriate approach for handling high-dimensional data that preserves key crown features while avoiding redundant learning during training. Forest conditions were found to affect CNN models in varied ways. For each specific condition, the requirements for delivering a desired ITDCD accuracy differed. Since deep learning-based object detection is still an emerging area, new models and approaches are continuously proposed and tested for ITDCD tasks. As a result, it is unfair to state that one type of model, either one-stage or two-stage, would universally outperform the others. However, as ITDCD focuses on a specific type of object, trees, large CNN models that are designed for comprehensive object detection may not always be required. Instead, a smaller CNN model that has higher learning efficiency with a limited number of tree samples may be preferable. More detailed analysis of how CNNs extract tree features (e.g. analysing feature and activation maps) would help to interpret trained models and guide future model selection for ITDCD tasks.

Appendix

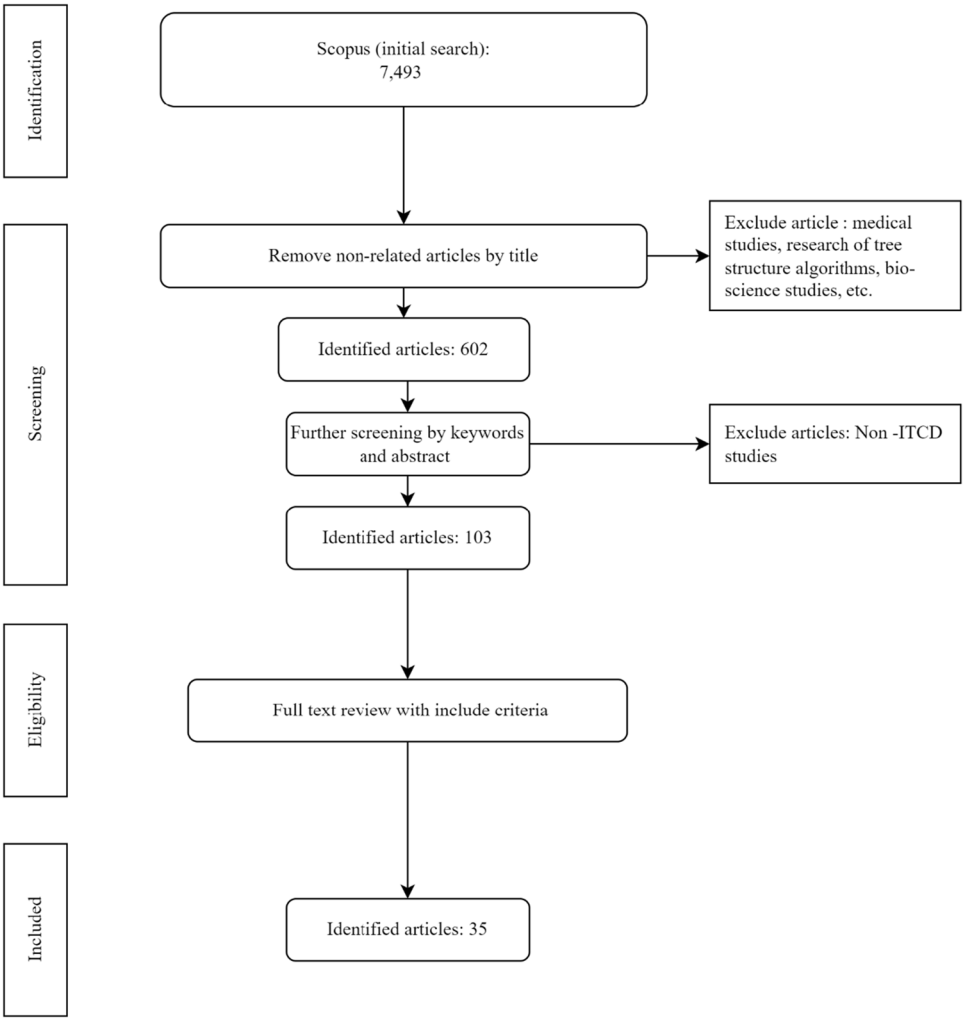

Fig. 3 Selection process of reviewed articles using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework. The chart structure was created by Moher et al. (2009)

Table 7 Data source of reviewed articles

Journal	Count of journal
Remote Sensing	11 [14••, 28, 30••, 39, 42, 47, 55, 57, 58, 62, 68]
ISPRS Journal of Photogrammetry and Remote Sensing	3 [48•, 59, 66]
IEEE Access	2 [38, 75]
Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering	2 [49, 67]
Computers and Electronics in Agriculture	2 [29, 43]
Others	15 [27, 40, 41, 50, 51, 56, 63, 64, 69•, 70–74, 76]
Agronomy *Engineering Applications of Artificial Intelligence* *Ecological Informatics* *Forest Ecology and Management* *Frontiers in Environmental Science* *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* *Journal of Forestry Research* *Journal of Sensors* *Journal of Spatial Science;Linye Kexue/Scientia Silvae Sinicae* *Methods in Ecology and Evolution* *Remote Sensing Letters* *Sensors* *Sensors (Switzerland)* *Sustainability (Switzerland)*	
Grand total	35

Table 8 Data synthesis items and descriptions

Data synthesis items		Description
ITDCD task		Individual tree detection (bounding box) or individual tree crown delineation (crown polygon)
Extracted tree class(es)		Tree class(es) extracted by CNN model. The classes could be tree species or tree categories (e.g. trees with different health conditions)
Dataset	Forest type	Natural forest, plantation forest, urban forest, mixed forest (multiple forest types)
	Extent of study area	Area used for training, classification, and validation (ha). For studies with multiple sites, a sum of areas was calculated
	Platform	UAV, satellite, or other aerial platforms
	Data type	Remote sensing data types used by CNNs, including RGB, multispectral, and lidar data
	Data specifications	Ground sample distance (GSD) for raster remote sensing data. For lidar data, GSD of derived raster surface was used. For studies that did not provide GSD, an estimate resolution was calculated based on camera specifications and flight heights
Training	CNN framework	CNN framework(s) used in a study
	Backbone	Backbone(s) used by CNN framework(s)
	Pre-training	Whether the backbone(s) was pre-trained and what dataset was used
	Samples (train/test/validation)	The sample numbers and ratio between training, test, and validation
	Training specs	Summary of configurations and techniques used in the training process
Accuracy assessment and output	Metrics	Accuracy metrics used for result assessment
	Result	Summary of accuracies and key results from a study
	Conclusion	Summary of key points of conclusions from a study
	Limitation	Summary of key points of limitations from a study and identify the potential opportunities for future research

Author Contribution Haotian Zhao carried out the literature review and drafted the manuscript. Justin Morgenroth, Grant Pearse and Jan Schindler provided critical feedback during the process and helped to finalize the manuscript. Grant Pearse and Jan Schindler also assisted in seeking financial support for this research.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This work was funded by the New Zealand Ministry of Business, Innovation and Employment under contract C09X1923 (Catalyst: Strategic Fund).

Data Availability The data used in this article is available upon request. Please contact the corresponding author for access to the data.

Compliance with Ethical Standards

Conflict of Interest The authors declare no competing of interests.

Human and Animal Rights and Informed This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. FAO UNEP. Forests, biodiversity and people. The State of the World's Forests (SOFO). Rome, Italy: FAO and UNEP; 2020. #214. https://doi.org/10.4060/ca8642en.
2. Fujimoto A, Haga C, Matsui T, Machimura T, Hayashi K, Sugita S, et al. An end to end process development for UAV-SfM based forest monitoring: Individual tree detection, species classification and carbon dynamics simulation. Forests. 2019;10(8). https://doi.org/10.3390/f10080680.
3. Saarinen N, Vastaranta M, Näsi R, Rosnell T, Hakala T, Honkavaara E, et al. Assessing biodiversity in boreal forests with UAV-based photogrammetric point clouds and hyperspectral imaging. Remote Sens. 2018;10(2):338. https://doi.org/10.3390/rs10020338.
4. Kimball LL, Wiseman PE, Day SD, Munsell JF. Use of urban tree canopy assessments by localities in the Chesapeake Bay Watershed. Cities and the Environment. 2014;7(2):9.
5. Berland A, Shiflett SA, Shuster WD, Garmestani AS, Goddard HC, Herrmann DL, et al. The role of trees in urban stormwater management. Landscape and Urban Plan. 2017;162:167–77. https://doi.org/10.1016/j.landurbplan.2017.02.017.
6. Brūmelis G, Dauškane I, Elferts D, Strode L, Krama T, Krams IJF. Estimates of tree canopy closure and basal area as proxies for tree crown volume at a stand scale. Forests. 2020;11(11):1180. https://doi.org/10.3390/f11111180.
7. Livesley S, McPherson EG, Calfapietra C. The urban forest and ecosystem services: impact on urban water, heat, and pollution cycles at the tree, street, and city scale. J Environ Qual. 2016;45:119–24. https://doi.org/10.2134/jeq2015.11.0567.
8. Shendryk I, Broich M, Tulbure MG, McGrath A, Keith D, Alexandrov SVJRSOE. Mapping individual tree health using full-waveform airborne laser scans and imaging spectroscopy A case study for a floodplain eucalypt forest. Remote Sens Environ. 2016;187:202–17. https://doi.org/10.1016/j.rse.2016.10.014.
9. Wang L, Gong P, Biging GSJPE, Sensing R. Individual tree-crown delineation and treetop detection in high-spatial-resolution aerial imagery. Photogramm Eng Rem S. 2004;70(3):351–7. https://doi.org/10.14358/PERS.70.3.35.
10. Alonzo M, Bookhagen B, Roberts DA. Urban tree species mapping using hyperspectral and lidar data fusion. Remote Sens Environ. 2014;148:70–83. https://doi.org/10.1016/j.rse.2014.03.018.
11. Murtha P, Fournier R. Varying reflectance patterns influence photo interpretation of dead tree crowns. Can J Remote Sens. 1992;18(3):167–73.
12. Röder M, Latifi H, Hill S, Wild J, Svoboda M, Brůna J, et al. Application of optical unmanned aerial vehicle-based imagery for the inventory of natural regeneration and standing deadwood in post-disturbed spruce forests. Int J Remote Sens. 2018;39(15–16):5288–309. https://doi.org/10.1080/01431161.2018.1441568.
13. St-Onge B, Grandin S. Estimating the height and basal area at individual tree and plot levels in Canadian subarctic lichen woodlands using stereo worldview-3 images. Remote Sens. 2019;11(3). https://doi.org/10.3390/rs11030248.
14. •• Braga JRG, Peripato V, Dalagnol R, Ferreira MP, Tarabalka Y, Aragão LEOC, et al. Tree crown delineation algorithm based on a convolutional neural network. Remote Sens. 2020;12(8). https://doi.org/10.3390/RS12081288.**This study proposed a simple copy-paste approach to create synthetic samples for ITDCD in natural forest. The method largely improved the data efficiency and has potential to be extented to other types of forest, where crown appearances are varied and manual sample collections are difficult.**
15. Pinz A. A computer vision system for the recognition of trees in aerial photographs. Nasa Conf P. 1991;3099:111-24.
16. Ke Y, Quackenbush LJ. A review of methods for automatic individual tree-crown detection and delineation from passive remote sensing. Int J Remote Sens. 2011;32(17):4725–47. https://doi.org/10.1080/01431161.2010.494184.
17. Yin D, Wang L. How to assess the accuracy of the individual tree-based forest inventory derived from remotely sensed data: a review. Int J Remote Sens. 2016;37(19):4521–53. https://doi.org/10.1080/01431161.2016.1214302.
18. Zhen Z, Quackenbush LJ, Zhang L. Trends in automatic individual tree crown detection and delineation-evolution of LiDAR data. Remote Sens. 2016;8(4). https://doi.org/10.3390/rs8040333.
19. Jing L, Hu B, Li J, Noland T, Guo H. Automated tree crown delineation from imagery based on morphological techniques. IOP C Ser Earth Env. 2014;17(1):012066. https://doi.org/10.1088/1755-1315/17/1/012066.
20 Jing L, Hu B, Li J, Noland T. Automated delineation of individual tree crowns from lidar data by multi-scale analysis and

segmentation. Photogramm Eng Rem S. 2012;78(12):1275–84. https://doi.org/10.14358/PERS.78.11.1275.
21. Qiu L, Jing L, Hu B, Li H, Tang YJRS. A new individual tree crown delineation method for high resolution multispectral imagery. Remote Sens. 2020;12(3):585. https://doi.org/10.3390/rs12030585.
22. Xu W, Deng S, Liang D, Cheng X. A crown morphology-based approach to individual tree detection in subtropical mixed broadleaf urban forests using UAV lidar data. Remote Sens. 2021;13(7). https://doi.org/10.3390/rs13071278.
23. Hoeser T, Kuenzer CJRS. Object detection and image segmentation with deep learning on Earth observation data: A review-part I: Evolution and recent trends. Remote Sens. 2020;12(10):1667. https://doi.org/10.3390/rs12101667.
24.•• Kattenborn T, Leitloff J, Schiefer F, Hinz SJIJOP, Sensing R. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS J Photogramm Remote Sens. 2021;173:24–49. **This review provides a comprehensive explanation of the concepts of CNN structures and techniques. It also provides a high-level view on the applications of CNN in vegetation analysis, which is a close domain to ITDCD**.
25. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
26. Xiao C, Qin R, Huang X. Treetop detection using convolutional neural networks trained through automatically generated pseudo labels. Int J Remote Sens. 2020;41(8):3010–30. https://doi.org/10.1080/01431161.2019.1698075.
27. Lou X, Huang Y, Fang L, Huang S, Gao H, Yang L, et al. Measuring loblolly pine crowns with drone imagery through deep learning. J For Res. 2021. https://doi.org/10.1007/s11676-021-01328-6.
28. Chadwick AJ, Goodbody TRH, Coops NC, Hervieux A, Bater CW, Martens LA, et al. Automatic delineation and height measurement of regenerating conifer crowns under leaf-off conditions using uav imagery. Remote Sens. 2020;12(24):1–26. https://doi.org/10.3390/rs12244104.
29. Wu J, Yang G, Yang H, Zhu Y, Li Z, Lei L, et al. Extracting apple tree crown information from remote imagery using deep learning. Comput Electron Agric. 2020;174. https://doi.org/10.1016/j.compag.2020.105504.
30.•• Weinstein BG, Marconi S, Bohlman S, Zare A, White E. Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. Remote Sens. 2019;11(11). https://doi.org/10.3390/rs11111309. **This paper proposed a semi-automated method to generate a large number of training samples with with minimal human labor required.. The method opens up a new direction for resolving data scarity in deep learning based ITDCD**.
31. Hoeser T, Bachofer F, Kuenzer CJRS. Object detection and image segmentation with deep learning on earth observation data: a review—Part II: Applications. Remote Sens. 2020;12(18):3053. https://doi.org/10.3390/rs12183053.
32. Boogaard FP, Rongen KSAH, Kootstra GW. Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging. Biosyst Eng. 2020;192:117–32. https://doi.org/10.1016/j.biosystemseng.2020.01.023.
33. Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: a meta-analysis and review. ISPRS J Photogramm Remote Sens. 2019;152:166–77. https://doi.org/10.1016/j.isprsjprs.2019.04.015.
34. Diez Y, Kentsch S, Fukuda M, Caceres MLL, Moritake K, Cabezas M. Deep learning in forestry using uav-acquired rgb data: a practical review. Remote Sens. 2021;13(14). https://doi.org/10.3390/rs13142837.
35. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group* t. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151(4):264–9. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.
36. Koirala A, Walsh KB, Wang Z, McCarthy C. Deep learning–method overview and review of use for fruit detection and yield estimation. Comput Electron Agr. 2019;162:219–34. https://doi.org/10.1016/j.compag.2019.04.017.
37 Everingham M, Van Gool L, Williams CK, Winn J, Zisserman AJIJOCV. The pascal visual object classes (voc) challenge. Int J Comput Vision. 2010;88(2):303–38.
38. Chiang CY, Barnes C, Angelov P, Jiang R. Deep learning-based automated forest health diagnosis from aerial images. IEEE Access. 2020;8:144064–76. https://doi.org/10.1109/ACCESS.2020.3012417.
39. Culman M, Delalieux S, Van Tricht K. Individual palm tree detection using deep learning on RGB imagery to support tree inventory. Remote Sens. 2020;12(21):1–31. https://doi.org/10.3390/rs12213476.
40. Ammar A, Koubaa A, Benjdira B. Deep-learning-based automated palm tree counting and geolocation in large farms from aerial geotagged images. Agronomy. 2021;11(8). https://doi.org/10.3390/agronomy11081458.
41. Park HG, Yun JP, Kim MY, Jeong SH. Multichannel object detection for detecting suspected trees with pine wilt disease using multispectral drone imagery. IEEE J Sel Top Appl Earth Obs Remote Sens. 2021;14:8350–8. https://doi.org/10.1109/JSTARS.2021.3102218.
42. Pleşoianu AI, Stupariu MS, Şandric I, Pătru-Stupariu I, Drăguţ L. Individual tree-crown detection and species classification in very high-resolution remote sensing imagery using a deep learning ensemble model. Remote Sens. 2020;12(15). https://doi.org/10.3390/RS12152426.
43. Xi X, Xia K, Yang Y, Du X, Feng H. Evaluation of dimensionality reduction methods for individual tree crown delineation using instance segmentation network and UAV multispectral imagery in urban forest. Comput Electron Agric. 2021;191. https://doi.org/10.1016/j.compag.2021.106506.
44. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. European conference on computer vision: Springer; 2014;740–55. https://doi.org/10.1007/978-3-319-10602-1_48.
45. Oksuz K, Cam BC, Akbas E, Kalkan S. Localization recall precision (LRP): a new performance metric for object detection. Proceedings of the European Conference on Computer Vision (ECCV) 2018;504–19.
46. Xu C, Morgenroth J, Manley B. Integrating data from discrete return airborne LiDAR and optical sensors to enhance the accuracy of forest description: a review. Curr For Rep. 2015;1:206–19. https://doi.org/10.1007/s40725-015-0019-3.
47. Pulido D, Salas J, Rös M, Puettmann K, Karaman S. Assessment of tree detection methods in multispectral aerial images. Remote Sens. 2020;12(15). https://doi.org/10.3390/RS12152379.
48.• Zheng J, Fu H, Li W, Wu W, Yu L, Yuan S, et al. Growing status observation for oil palm trees using unmanned aerial vehicle (UAV) images. ISPRS J Photogramm Remote Sens. 2021;173:95–121. https://doi.org/10.1016/j.isprsjprs.2021.01.008. **This study proposed a customized CNN structure to reduce the impact of imbalanced data, which is a common issue in ITDCD applications but was discussed very little in reviewed studies.**
49. Huang L, Wang Y, Xu Q, Liu Q. Recognition of abnormally discolored trees caused by pine wilt disease using YOLO algorithm and UAV images. Nongye Gongcheng Xuebao.

2021;37(14):197–203. https://doi.org/10.11975/j.issn.1002-6819.2021.14.022.
50. Paul A, Bhattacharyya S, Chakraborty D. Estimation of shade tree density in tea garden using remote sensing images and deep convolutional neural network. J Spat Sci. 2012;29:1–5. https://doi.org/10.1080/14498596.2021.2013966.
51. Emin M, Anwar E, Liu S, Emin B, Mamut M, Abdukeram A, et al. Target detection-based tree recognition in a spruce forest area with a high tree density—implications for estimating tree numbers. Sustainability. 2021;13(6). https://doi.org/10.3390/su13063279.
52. Maschler J, Atzberger C, Immitzer M. Individual tree crown segmentation and classification of 13 tree species using Airborne hyperspectral data. Remote Sens. 2018;10(8). https://doi.org/10.3390/rs10081218.
53. Naveed F, Hu B, Wang J, Hall GB. Individual tree crown delineation using multispectral LiDAR data. Sensors. 2019;19(24). https://doi.org/10.3390/s19245421.
54. Ozdarici-Ok A. Automatic detection and delineation of citrus trees from VHR satellite imagery. Int J Remote Sens. 2015;36(17):4275–96. https://doi.org/10.1080/01431161.2015.1079663.
55. Mo J, Lan Y, Yang D, Wen F, Qiu H, Chen X, et al. Deep learning-based instance segmentation method of litchi canopy from uav-acquired images. Remote Sens. 2021;13(19). https://doi.org/10.3390/rs13193919.
56. Safonova A, Guirado E, Maglinets Y, Alcaraz-Segura D, Tabik S. Olive tree biovolume from uav multi-resolution image segmentation with mask r-cnn. Sensors. 2021;21(5):1–17. https://doi.org/10.3390/s21051617.
57. Ampatzidis Y, Partel V. UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. Remote Sens. 2019;11(4). https://doi.org/10.3390/rs11040410.
58. Windrim L, Bryson M. Detection, segmentation, and model fitting of individual tree stems from airborne laser scanning of forests using deep learning. Remote Sens. 2020;12(9). https://doi.org/10.3390/RS12091469.
59. Hao Z, Lin L, Post CJ, Mikhailova EA, Li M, Chen Y, et al. Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (Mask R-CNN). ISPRS J Photogramm Remote Sens. 2021;178:112–23. https://doi.org/10.1016/j.isprsjprs.2021.06.003.
60. Zheng X, Wu X, Huan L, He W, Zhang H. A Gather-to-guide network for remote sensing semantic segmentation of rgb and auxiliary image. IEEE Trans Geosci Remote Sens. 2021;60:1–15. https://doi.org/10.1109/TGRS.2021.3103517.
61. Van Etten A. You only look twice: rapid multi-scale object detection in satellite imagery. arXiv preprint arXiv:180509512. 2018. https://doi.org/10.48550/arXiv.1805.09512.
62. Fromm M, Schubert M, Castilla G, Linke J, McDermid G. Automated detection of conifer seedlings in drone imagery using convolutional neural networks. Remote Sens. 2019;11(21). https://doi.org/10.3390/rs11212585.
63. Ocer NE, Kaplan G, Erdem F, KucukMatci D, Avdan U. Tree extraction from multi-scale UAV images using Mask R-CNN with FPN. Remote Sens Lett. 2020;11(9):847–56. https://doi.org/10.1080/2150704X.2020.1784491.
64. Weinstein BG, Marconi S, Aubry-Kientz M, Vincent G, Senyondo H, White EP. DeepForest: a Python package for RGB deep learning tree crown delineation. Methods Ecol Evol. 2020;11(12):1743–51. https://doi.org/10.1111/2041-210X.13472.
65. Lee AL, To CC, Lee AL, Li JJ, Chan RC. Model architecture and tile size selection for convolutional neural network training for non-small cell lung cancer detection on whole slide images. Inform Med Unlocked. 2022;28:100850. https://doi.org/10.1016/j.imu.2022.100850.
66. Gomez Selvaraj M, Vergara A, Montenegro F, Alonso Ruiz H, Safari N, Raymaekers D, et al. Detection of banana plants and their major diseases through aerial images and machine learning methods: a case study in DR Congo and Republic of Benin. ISPRS J Photogramm Remote Sens. 2020;169:110–24. https://doi.org/10.1016/j.isprsjprs.2020.08.025.
67. Ye Y, Shen B, Shen Y. Research on anti-shadow tree detection method based on generative adversarial network. Nongye Gongcheng Xuebao. 2021;37(10):118–26. https://doi.org/10.11975/j.issn.1002-6819.2021.10.014.
68. Morgenroth J, Östberg J, Van den Bosch CK, Nielsen AB, Hauer R, Sjöman H, et al. Urban tree diversity—Taking stock and looking ahead. Urban For Urban Gree. 2016;15:1–5. https://doi.org/10.1016/j.ufug.2015.11.003.
69.• Zamboni P, Junior JM, Silva JA, Miyoshi GT, Matsubara ET, Nogueira K, et al. Benchmarking anchor-based and anchor-free state-of-the-art deep learning methods for individual tree detection in rgb high-resolution images. Remote Sens. 2021;13(13). https://doi.org/10.3390/rs13132482. **This study compared 21 CNN models from three major categories for an urban ITDCD task. Compared with one-stage and two-stage models, which are widely used in other ITDCD studies, their result shows the potential of anchor-based models**.
70. Xia K, Wang H, Yang Y, Du X, Feng H. Automatic detection and parameter estimation of Ginkgo biloba in urban environment based on RGB images. Journal of Sensors. 2021;2021. https://doi.org/10.1155/2021/6668934.
71. Zhou Y, Liu W, Luo Y, Zong S. Small object detection for infected trees based on the deep learning method. Linye Kexue. 2021;57(3):98–107. https://doi.org/10.11707/j.1001-7488.20210310.
72. Ferreira MP, Almeida DRAD, Papa DDA, Minervino JBS, Veras HFP, Formighieri A, et al. Individual tree detection and species classification of Amazonian palms using UAV images and deep learning. For Ecol Manage. 2020;475. https://doi.org/10.1016/j.foreco.2020.118397.
73. Tong P, Han P, Li S, Li N, Bu S, Li Q, et al. Counting trees with point-wise supervised segmentation network. Eng Appl Artif Intell. 2021;100. https://doi.org/10.1016/j.engappai.2021.104172.
74. Li F, Liu Z, Shen W, Wang Y, Wang Y, Ge C, et al. A remote sensing and airborne edge-computing based detection system for pine wilt disease. IEEE Access. 2021. https://doi.org/10.1109/ACCESS.2021.3073929.
75. Dos Santos AA, Marcato Junior J, Araújo MS, Di Martini DR, Tetila EC, Siqueira HL, et al. Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVS. Sensors. 2019;19(16). https://doi.org/10.3390/s19163595.
76. Weinstein BG, Marconi S, Bohlman SA, Zare A, White EP. Cross-site learning in deep learning RGB tree crown detection. Ecol Informatics. 2020;56. https://doi.org/10.1016/j.ecoinf.2020.101061.
77. Zheng Y, Wu G. Single shot MultiBox detector for urban plantation single tree detection and location with high-resolution remote sensing imagery. Front Environ Sci. 2021;9. https://doi.org/10.3389/fenvs.2021.755587.
78. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014. https://doi.org/10.48550/arXiv.1409.1556.
79. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of IEEE Conf Comput Vis Pattern Recognit. 2015;1–9.

80. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of IEEE Conf Comput Vis Pattern Recognit. 2016. pp. 779–88.
81. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;770–8. https://doi.org/10.1109/CVPR.2016.90.
82. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. International conference on machine learning: PMLR; In: Proceedings of Int Conf Mach Learn. 2019;6105–14.
83. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2117–25.
84. Mahmud MS, Zahid A, Das AK, Muzammil M, Khan MU. A systematic literature review on deep learning applications for precision cattle farming. Comput Electron Agric. 2021;187:106313. https://doi.org/10.1016/j.compag.2021.106313.
85. Zhao ZQ, Zheng P, Xu ST, Wu X. Object detection with deep learning: a review. IEEE T Neur Net Lear. 2019;30(11):3212–32. https://doi.org/10.1109/TNNLS.2018.2876865.
86. Wang Y, Albrecht CM, Braham NAA, Mou L, Zhu XX. Self-supervised learning in remote sensing: a review. arXiv preprint arXiv:220613188. 2022. https://doi.org/10.48550/arXiv.2206.13188.
87. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imag Radiat On. 2021;65(5):545–63. https://doi.org/10.1111/1754-9485.13261.
88. Xiao Y, Tian Z, Yu J, Zhang Y, Liu S, Du S, et al. A review of object detection based on deep learning. Multimedia Tools Appl. 2020;79(33–34):23729–91. https://doi.org/10.1007/s11042-020-08976-6.
89. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. IEEE Conf Comput Vis Pattern Recognit. 2009;248–55. https://doi.org/10.48550/arXiv.2206.13188.
90. Soviany P, Ionescu RT. Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In: 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC): IEEE; 2018;209–14. https://doi.org/10.1109/SYNASC.2018.0004.
91. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE I Conf Comp Vis.2017;2980–8.
92. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: Single shot multibox detector. European conference on computer vision: Springer; 2016; 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
93. Ge Z, Liu S, Wang F, Li Z, Sun J. Yolox: exceeding yolo series in 2021. arXiv preprint arXiv:210708430. 2021. https://doi.org/10.48550/arXiv.2107.08430.
94. Wang CY, Bochkovskiy A, Liao HYM. Scaled-yolov4: scaling cross stage partial network. In: Proceedings 2021 IEEE Conf Comp Vis Pattern Recognit, IEEE Comput Soc. 13024–33. https://doi.org/10.1109/CVPR46437.2021.01283.
95. Ward D, Moghadam P. Scalable learning for bridging the species gap in image-based plant phenotyping. Comput Vision Image Understanding. 2020;197–198. https://doi.org/10.1016/j.cviu.2020.103009.
96. Wang C, Liu B, Liu L, Zhu Y, Hou J, Liu P, et al. A review of deep learning used in the hyperspectral image analysis for agriculture. Artif Intell Rev. 2021;54(7):5205–53. https://doi.org/10.1007/s10462-021-10018-y.
97. Oksuz K, Cam BC, Kalkan S, Akbas E. Imbalance problems in object detection: a review. IEEE Trans Pattern Anal Mach Intell. 2020;43(10):3388–415. https://doi.org/10.1109/TPAMI.2020.2981890.
98. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. Journal of big data. 2019;6(1):1–48.
99. Ophoff T, Van Beeck K, Goedemé TJS. Exploring RGB+ depth fusion for real-time object detection. 2019;19(4):866. https://doi.org/10.3390/s19040866.
100. Yeong DJ, Velasco-Hernandez G, Barry J, Walsh JJS. Sensor and sensor fusion technology in autonomous vehicles: a review. 2021;21(6):2140. https://doi.org/10.3390/s21062140.
101. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6(1). https://doi.org/10.1186/s40537-019-0192-5.
102. Bissoto A, Valle E, Avila S. Gan-based data augmentation and anonymization for skin-lesion analysis: a critical review. In: Proceedings 2021 IEEE Conf Comp Vis Pattern Recognit, IEEE Comput Soc 1847–56.
103. Bi L, Hu G. Improving image-based plant disease classification with generative adversarial network under limited training set. Front Plant Sci. 2020;11. https://doi.org/10.3389/fpls.2020.583438.
104. Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, et al. GAN-based synthetic brain MR image generation. 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018): IEEE; 2018;734–8. https://doi.org/10.1109/ISBI.2018.8363678.
105. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan HJN. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing. 2018;321:321–31. https://doi.org/10.1016/j.neucom.2018.09.013.
106. Nguyen N-D, Do T, Ngo TD, Le D-D. An evaluation of deep learning methods for small object detection. Journal of Electrical and Computer Engineering. 2020;2020.
107. Justus D, Brennan J, Bonner S, McGough AS. Predicting the computational cost of deep learning models. IEEE Int Conf Big Data. 2018;3873–82. https://doi.org/10.1109/BigData.2018.8622396.
108.• Yu T, Zhu H. Hyper-parameter optimization: a review of algorithms and applications. arXiv preprint arXiv:200305689. 2020. https://doi.org/10.48550/arXiv.2003.05689.**The study describes the impacts of parameter optimisation and provides introductions around several techniques. The findings are valuable for further improving the performance of CNN models on ITDCD tasks**.
109. Condés S, Sterba HJFE, Management. Derivation of compatible crown width equations for some important tree species of Spain. Forest Ecol Manag.2005;217(2–3):203–18. https://doi.org/10.1016/j.foreco.2005.06.002.