ISPRS Journal of Photogrammetry and Remote Sensing 232 (2026) 974–999

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Empowering tree-scale monitoring over large areas: Individual tree delineation from high-resolution imagery

Xinlian Liang [a,*], Yinrui Wang [a], Jun Pan [a,b], Janne Heiskanen [c,d], Ningning Wang [e], Siyu Wu [f], Ilja Vuorinne [c,g], Jiaojiao Tian [h,i], Jonas Troles [j], Myriam Cloutier [k], Stefano Puliti [l], Aishwarya Chandrasekaran [m], James Ball [n], Xiangcheng Mi [e], Guochun Shen [f], Kun Song [f], Guofan Shao [o], Rasmus Astrup [l], Yunsheng Wang [p], Petri Pellikka [c,a], Mi Wang [a,b], Jianya Gong [q]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China
[b] Oriental Space Port Research Institute, Yantai 265100, China
[c] Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, 00014 Helsinki, Finland
[d] Finnish Meteorological Institute, P.O. Box 503, 00101 Helsinki, Finland
[e] Zhejiang Qianjiangyuan Forest Biodiversity National Observation and Research Station, Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[f] Zhejiang Tiantong Forest Ecosystem National Observation and Research Station, School of Ecological and Environmental Sciences, East China Normal University, Shanghai 200241, China
[g] Institute for Atmospheric and Earth System Research, University of Helsinki, P.O. Box 64, 00014 Helsinki, Finland
[h] Remote Sensing Technology Institute, German Aerospace Center, Wesslin 82234, Germany
[i] Institute of Forest Management, School of Life Sciences Weihenstephan, Technical University of Munich, Freising 85354, Germany
[j] Cognitive Systems Group, University of Bamberg, Bamberg 96050, Germany
[k] Department of Forest & Conservation Sciences, Faculty of Forestry, The University of British Columbia, Canada
[l] Norwegian Institute of Bioeconomy Research, PO Box 115, 1430 Aas, Norway
[m] Department of Environment and Society, Utah State University, Logan, UT 84322, USA
[n] Conservation Research Institute, Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK
[o] Department of Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907, USA
[p] Department of Remote Sensing and Photogrammetry, Finnish Geospatial Research Institute (National Land Survey of Finland), Vuorimiehentie 5, 02150 Espoo, Finland
[q] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Forest
Individual tree delineation
Remote Sensing
Imagery
Deep learning
Earth observation
Close-range

ABSTRACT

Accurate individual tree delineation (ITD) is essential for forest monitoring, biodiversity assessment, and ecological modeling. While remote sensing (RS) has significantly advanced forest ITD, challenges persist, especially in complex forest environments. The use of imagery data is compelling given the rapid increase in available high-resolution aerial and satellite imagery data, the increasing need for image-based analysis where reliable 3D data are unavailable, the widening gap between data supply and processing capabilities, and the limited validation of state-of-the-art (SOTA) methods across diverse real-world conditions. This study aims to advance ITD research by evaluating SOTA instance segmentation approaches, including both recently developed and established methods. The analysis evaluates ITD algorithm performance using the largest forest instance-segmentation imagery dataset to date and standardized evaluation protocols. This study identifies key factors affecting accuracy, reveals remaining challenges, and outlines future research directions. Findings in this study reveal that ITD accuracy is heavily influenced by image resolution, forest structure, and method design. Findings also reveal that, while algorithm innovations remain important, robustness and transferability that ensure

* Corresponding author.
E-mail addresses: xinlian.liang@whu.edu.cn (X. Liang), wangyinrui@whu.edu.cn (Y. Wang), panjun1215@whu.edu.cn (J. Pan), janne.heiskanen@helsinki.fi (J. Heiskanen), wangningning@ibcas.ac.cn (N. Wang), 18289665799@163.com (S. Wu), ilja.vuorinne@helsinki.fi (I. Vuorinne), jiaojiao.tian@dlr.de (J. Tian), jonas.troles@uni-bamberg.de (J. Troles), myriamcl@student.ubc.ca (M. Cloutier), stefano.puliti@nibio.no (S. Puliti), aish.chandrasekaran@usu.edu (A. Chandrasekaran), jgcb3@cam.ac.uk (J. Ball), mixiangcheng@ibcas.ac.cn (X. Mi), gcshen@des.ecnu.edu.cn (G. Shen), ksong@des.ecnu.edu.cn (K. Song), shao@purdue.edu (G. Shao), rasmus.astrup@nibio.no (R. Astrup), yunsheng.wang@nls.fi (Y. Wang), petri.pellikka@helsinki.fi (P. Pellikka), wangmi@whu.edu.cn (M. Wang), gongjy@whu.edu.cn (J. Gong).

https://doi.org/10.1016/j.isprsjprs.2025.12.022
Received 10 February 2025; Received in revised form 22 December 2025; Accepted 22 December 2025
Available online 20 January 2026

generalization across diverse environments are what differentiate method performances. In addition, this study highlights that commonly used evaluation metrics may fail to adequately capture precise performance in specific applications, e.g., individual-tree-crown segmentation in this study. Assessment reliability can be strengthened through the adoption of stricter criteria. Future research should focus on expanding datasets, refining evaluation protocols, and developing adaptive models capable of handling varying canopy structures. These advancements will enhance ITD scalability and reliability, contributing to more effective forest research and management at a global scale.

1. Introduction

The forest canopy refers to the uppermost layer of a forest, formed by the aggregated crowns of the upper layer trees and their associated organisms. It represents a functional and structural continuum shaped by interactions between crowns and non-tree components such as epiphytes. The forest canopy is structurally complex and ecologically critical (Lowman and Rinker, 2004). It functions as a primary engine of forest ecosystem productivity, where high light availability drives efficient photosynthesis. Beyond this, the canopy regulates local environmental conditions such as solar radiation, humidity, and temperature, therefore directly influences overall ecosystem functionality, from tree growth, resilience, to biodiversity.

Individual tree crowns are the structural and fundamental units of the forest canopy. Consequently, crown-level information is critical for understanding canopy processes. For example, the crown dimensions (e.g., length, width, and volume) and configuration (vertical stratification and horizontal distribution) directly govern light penetration, energy distribution, and humidity gradients; the crown architecture (e.g., conical, spreading) shapes canopy heterogeneity, which supports niche differentiation for species; the crown-level traits (e.g., branching patterns and leaf density) create complex three-dimensional (3D) microhabitats that support epiphytes, insects, birds, and mammals, which are fundamental to sustaining biodiversity. While studies of forest canopy are central to advancing ecological understanding (Spies, 1998), accurate tree-scale assessment of crown characteristics remains a significant challenge.

Tree crowns are conventionally visually observed *in situ* from the ground at a local or a stand scale. Due to the lack of effective tools and/or high costs of data acquisition, capturing true structure of the canopies with conventional field measurements is constrained by various practical difficulties (Goodman et al., 2014). Therefore, tree-crown attributes (e.g., the shape, volume, architecture, and complexity) have to be represented by simplified features, if quantitatively measured at all, and the canopy representations are often subject to significant uncertainty (Liang et al., 2022b).

LiDAR (Light Detection and Ranging) advances tree-scale studies by canopy penetration and direct 3D structural measurements. Individual tree delineation (ITD), also known as individual tree segmentation, algorithms quantify individual trees, either through LiDAR produced canopy height models (CHMs) (Kaartinen et al., 2012; Vauhkonen et al., 2011) or point cloud (Hyyppa et al., 2001; Wang et al., 2016). Beyond the tree locations, ITD often provides individual tree masks that further facilitate the characteristics assessment at the tree scale, such as branch structures (Pyörälä et al., 2018; Harikumar et al., 2022; Cai et al., 2024; Qi and Liang, 2024) and crown shapes (Liang et al., 2024a; Zhang et al., 2024). However, the coverage of high-density LiDAR operations is often limited due to hardware and cost constraints, restricting its capability to provide wall-to-wall assessments over large areas.

Recent advancements have taken two approaches to enhance tree-scale crown characterization. The first focuses on improving *in situ* tree-wise measurements by increasing the automation and intelligence of field investigations, e.g., (Hyyppä et al., 2020a, 2020b; Wang et al., 2021; Liang et al., 2024b). The other approach investigates ITD-based inventories directly from high spatial resolution (e.g., $<$0.5 m) aerial and earth observation (EO) data. Considering the higher accessibility and the lower costs, the EO-imagery based ITD started to gain attention as a complementary solution to *in situ* close-range sensing technologies for tree-scale characterizations.

Despite lacking height information (i.e., the third spatial dimension), high-resolution images provide detailed texture and spectral features of tree crowns with wall-to-wall coverage. High-resolution images can be collected from various platforms. Aerial platforms such as airplanes can capture centimeter(cm)-level resolution data over large areas. However, the high operational cost constrains their temporal resolution and utilization in monitoring dynamic processes. Drones or unmanned aerial vehicles (UAVs) have in recent years emerged as prevalent data sources for high-resolution aerial observations. While drones significantly lower the cost and logistics barrier in data acquisition, their inherent limitations in flight durance result in confined spatial coverage, making them ideal for site-specific studies rather than large-area mapping.

Meanwhile, high-resolution satellite EO images have become increasingly accessible through various satellites, such as WorldView 1–4, GeoEye, Pléiades, and Pléiades Neo, with improved resolution (e.g., sub-meter), frequency, and coverage. Particularly, satellites' high revisit frequency enables long-term time series with high spatial and temporal resolutions over large areas, offering advantages in the temporal domain that are rarely comparable by other platforms. This makes scalable, repeated, landscape observation operationally feasible, which results in a paradigm shift that operationalizes continuous, fine-grained monitoring of ecosystem dynamics across entire landscapes and closes the gap between localized detail and regional context. Yet, the capability of pure-imagery-based ITD approaches has not been adequately investigated, to leverage the potential of high-resolution aerial and satellite observations.

It is worth noting that the ITD from imagery fundamentally differs from that based on 3D data such as LiDAR and structure from motion (SfM), even though the tasks are conceptually similar, particularly when a canopy height model (CHM) is used as input. CHM is a 2D raster representation of 3D point cloud data, resembling an orthophoto of the forest canopy but encoding canopy height information instead of spectral reflectance. The depth information significantly shapes the delineation process and impacts associated accuracy. With the aid of canopy height information, CHM-based ITD is straightforward. Individual treetops such as the highest point of a crown can easily be identified by detecting local height maxima. Tree crown boundaries can then be delineated using image segmentation algorithms, such as region growing or morphological watershed, by allocating CHM pixels to the detected treetops based on proximity and height gradients (Hyyppa et al., 2001; Kaartinen et al., 2012). Though sharing similar segmentation ideas as those applied to imagery data, most algorithms that support CHM-based ITD cannot be directly applied to 2D optical images due to the absence of canopy height information.

ITD using high-resolution imagery relies on spectral and textural features, e.g., vegetation indices, crown edges and/or gaps between crowns, and texture metrics like GLCM (Gray-Level Cooccurrence Matrix). Conventional image processing approaches, such as object-based image analysis, edge detection, and morphological operations, require predefined rulesets based on crown size, circularity, compactness, and other factors (Kotaridis and Lazaridou, 2021). Examples can be found in the valley following (Gougeon, 1995) and watershed (Wang, 2003) methods applied on the aerial grey images. These methods rely on clear

gaps or boundaries between crowns, and are therefore typically constrained to low-density forests, homogeneous plantations, or coniferous forests (Jing et al., 2012; Safonova et al., 2021; Wang, 2010). The robustness and generalization of these machine learning (ML) methods remain unproven.

Compared to conventional image processing methods, deep learning (DL) has driven ITD approaches delivering improved performance, which is similar to what was observed in tree-scale species classification (Chen et al., 2024). Various instance segmentation neural networks (NNs) have been applied to image-based ITD tasks. Mask R-CNN (He et al., 2017) is one of the most widely used architectures, often implemented using imagery alone (Gan et al., 2023; Sani-Mohammed et al., 2022) or a fusion of RGB imagery and CHM (Xie et al., 2024). Improved versions of Mask R-CNN have also been explored, such as adding a Transformer-based contextual aggregation module to enhance texture information extraction and distinguish subtle canopy texture differences (Zhu et al., 2024). A semi-supervised learning scheme for Mask R-CNN training was introduced to reduce the workload of individual tree crown (ITC) labeling, using airborne RGB and CHM data (Dersch et al., 2024). Other commonly used NNs in the image-based ITD include Cascade Mask R-CNN (Chen et al., 2019) that had been applied to airborne RGB images in urban areas (Sun et al., 2022), and BlendMask (Chen et al., 2020) that was reported to outperform Mask R-CNN on UAV imagery.

This study aims to (1) reveal the performances of state-of-the-art (SOTA) methods for image-based ITD, (2) establish the performance baseline of image-based ITD through extensive testing and benchmarking, especially in challenging scenarios, (3) examine key challenges and impact factors for the algorithm performances, and (4) prospect the possible solutions for future advancements.

To resemble the general aerial and satellite image collection scenarios, this study focuses on the ITD methods that solely depend on non-overlapped imagery datasets, with no reliance on additional auxiliary data sources such as LiDAR. This is due to four main reasons: (1) interest in applying high-resolution aerial/satellite data in fine-scale applications such as individual tree studies are high, as such imagery is quickly accumulating and becoming widely available; (2) the need to apply imagery data as the main data source without the aid of 3D information are widespread due to the limited availability of reliable 3D data in many practical fine-scale applications, yet their potential has rarely been explored; (3) algorithm development lags behind the data supply, where the current studies concentrated on a few existing standard models, resulting in a gap between the huge amount of aerial/satellite imagery data and the available processing capabilities; and (4) the applicability of SOTA image-based methods has not been clarified, as existing experimental tests have been limited to small areas (e.g., mostly only one study site in reported studies).

The instance-segmentation methodology adopted in this study represents SOTA solutions and comprises four main components: (1) SOTA methods, including Cascade Mask R-CNN (He et al., 2017), Hybrid Task Cascade (HTC) (Chen et al., 2019), and Mask DINO (Li et al., 2023), which took the ResNet-50 (He et al., 2016) as backbone and were trained using the dataset in this study; (2) a general-purpose large model, Segment Anything (SAM) (Kirillov et al., 2023), which was pre-trained by a broad dataset and implemented by the zero-shot inference for ITD; (3) an ML ITD method based on the marker-controlled watershed, which serves as an example to better understand the performance of ML- and DL-based methods; (4) the top 6 ranking methods from the recent international ITD contest. Thus, this study comprehensively investigates the SOTA of existing ITD methodologies.

2. Materials and experiment

The experiment setups focus on assessing the applicability and transferability of SOTA methods across diverse forest types and data conditions, in order to reveal the SOTA of the image-based ITD, set a baseline for further research, and promote research in this field.

2.1. Dataset and reference information

Supervised approaches, both ML and DL methods, are data-driven and depend on high-quality datasets with extensive annotations. The datasets used in the contest and this study, comprising a significant amount of annotated ITD reference images, are the result of a collaborative effort from numerous contributors worldwide. The datasets include those that have been published before the contest and those newly released in this contest. The ITC annotations of all datasets were newly labeled or revised in the contest.

The dataset covers 11 study sites across 9 countries including Australia, Canada, China, Germany, Kenya, Malaysia, Norway, Panama, and United States, spanning tropical, subtropical, and temperate climate zones. The dataset represents diverse forest types, such as tropical rainforests, montane forests, moist forests, and savanna woodlands; subtropical evergreen broadleaf and mixed forests; and temperate broadleaf and mixed forests. In addition, the dataset also includes both natural and urban forests.

High-resolution images were acquired from the study sites using airplane or UAV platforms, with resolutions ranging from 2 cm to 10 cm. The original images were then cropped into individual image tiles with a size of 1024 × 1024 pixels for further processing. Table 1 lists detailed information about the study sites and the conditions of their corresponding dataset.

As the reference for the training and evaluation, 600,000 manually annotated ITC masks were provided for over 11,000 images. Annotations of the visible dominant and co-dominant ITC on images were conducted through visual interpretation.

Previously existing ITC annotations in study site 1 (Cloutier et al., 2024), 2 (Ball et al., 2023), 3 (Lee et al., 2023), 7 (Chandrasekaran et al., 2024), 10 (Troles et al., 2024), and 11 (Jansen et al., 2023) have common problems such as missing labels, incorrect boundaries, and merged ITCs. Such annotations were carefully revisited, verified, and revised, and missing labels were manually added. The original ITC annotations of study site 9 (Puliti and Astrup, 2022) are bounding boxes. The ITC boundaries of this site are newly labeled. Study site 4, 5, 6, and 8 do not have ITC annotations before the contest, and ITCs were manually annotated from scratch.

All annotations were manually delineated by trained staffs, and quality was checked, compared, and assessed by multiple forestry experts across each dataset, to ensure accuracy. Several iterative modifications and assessments were executed. The annotated ITC masks are organized by the standardized MS COCO data format (Lin et al., 2014).

The annotation process is time-intensive and requires substantial resources; however, it produces comprehensive references for both training and evaluation. Although a gap between reference and ground truth in reality is inevitable, e.g., mislabeling of invisible suppressed small trees, and confusion between adjacent crowns with similar texture, the reference represents annotations that are as accurate as possible through careful human interpretation of images.

The dataset is openly available. . The training and validation sets include both images and labels for ITD model development. The testing set includes only images. The performance on the testing set can be evaluated by submitting the ITD results to the online benchmark platform, where the ranking and scores are displayed on the leaderboard. The link of the benchmark platform is https://www.codabench.org/competitions/12668/?secret_key=e73a13fa-9245-4bd0-8ba9-64b761fddc66.

Fig. 1 illustrates the spatial distribution of the study sites along with sample images and corresponding references. The dataset's quantity, quality, and diversity set a new SOTA for image-based ITD studies.

2.2. Example code

A set of DL-based example code for model training, evaluation, and inference was developed and provided to all contest participants to help

Table 1
The detailed information of the datasets.

Type			DatasetID	Country	Resolution (cm)	New dataset	Images quantity		
Land cover	Climate	Forest					Training	Validation	Testing
Natural	Temperate	broadleaf	7	America	5.0	–	184	–	100
		mixed	1	Canada	2.0	–	1691	–	–
		mixed	9	Norway	2.0–7.0	–	206	–	100
		mixed	10	German	2.0	–	–	–	468
	Tropical	rainforest	2	Malaysia	10.0	–	331	–	–
		moist	3	Panama	4.0	–	1200	275	600
		montane	8	Kenya	10.0	√	300	–	200
		savanna woodland	11	Australia	2.0	–	–	–	200
	Sub-tropical	evergreen broad-leaf	4	China	10.0	√	400	100	200
		evergreen broadleaf	5	China	2.0	√	1721	441	786
Urban		mixed	6	China	3.0	√	1234	–	346

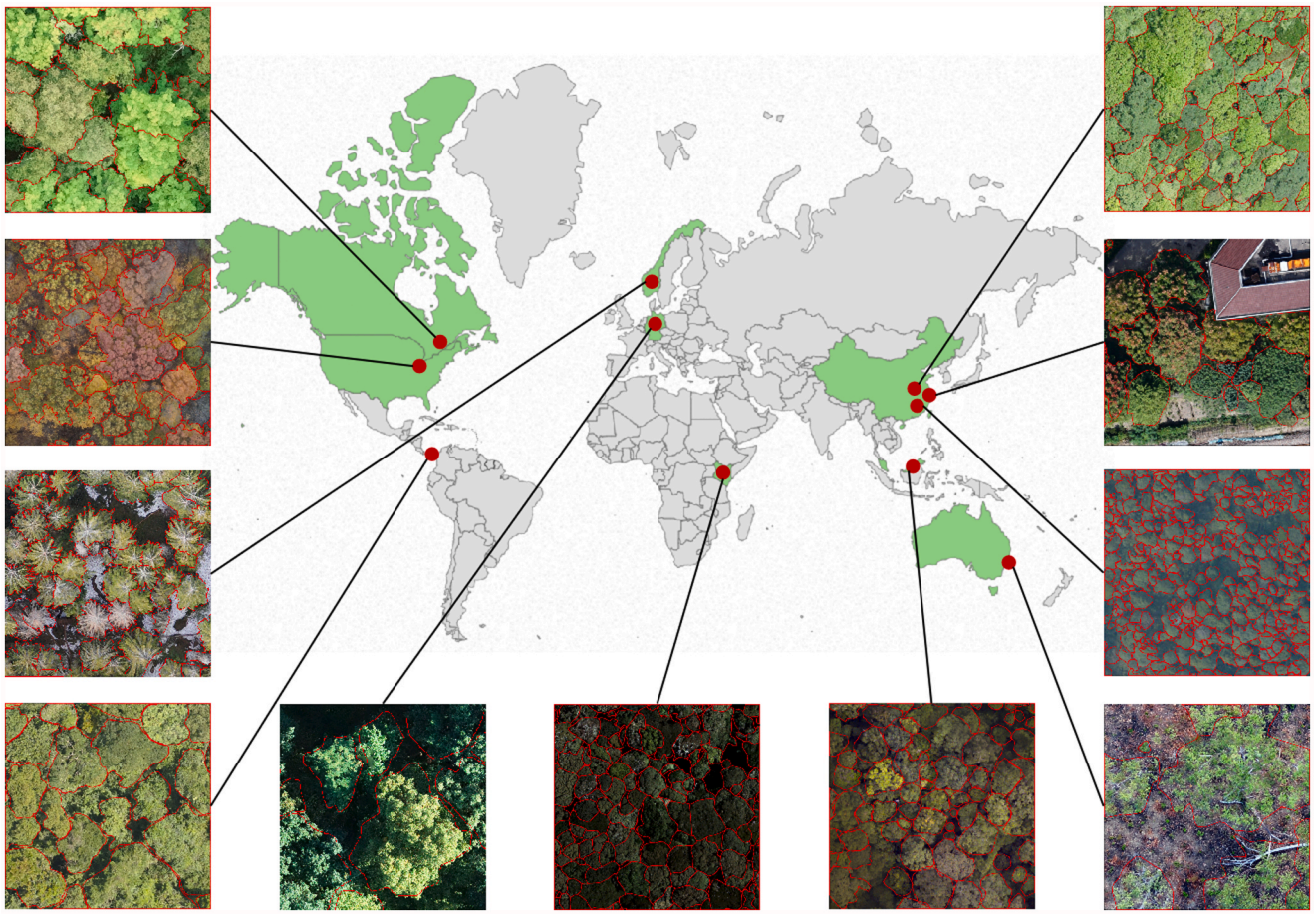

Fig. 1. The distribution of study sites.

them understand the contest setup, tasks, and evaluation procedures, as well as to accelerate the development of their own methods. The sample code employed the Mask R-CNN (He et al., 2017) instance segmentation network for ITD and was developed with reference to detectron2 introduced in (Wu et al., 2019). The example code is openly accessible from the address, https://github.com/MSpace-WHU/ITD_Contest_2024_sample_code.

2.3. The international individual-tree-delineation contest

Some of the methods investigated in this study are based on the first international contest on methods for image-based ITD 2024, sponsored by the International Society of Photogrammetry and Remote Sensing (ISPRS). This section briefly introduces the aims and implementations of the contest.

2.3.1. Overview

The international image-based ITD contest 2024 aims to promote the research of ITD using high-resolution earth observation imagery. The contest was open to all who were interested. The participants developed and tested their methods on standardized datasets and the results were benchmarked using a standardized evaluation framework. The results of the top teams were validated through submitted Docker files. The inference codes were implemented on the testing dataset, and the outcomes were validated by the organizer in order to verify the reported results in the test stages.

All conventional ML and DL approaches were welcomed in the contest, though the example code is a DL approach. However, according

to the final results in the contest, all top-listed teams that provided their results used DL approaches, which reflects the SOTA of algorithm development in this field.

Over 40 teams from 13 countries joined the contest, including universities, companies, and independent researchers. The background and the implementation of the contest were preliminarily reported in (Liang et al., 2024) as a conference presentation. This is the first research work that thoroughly introduces and analyzes the outcomes of the contest and reveals the insights learnt from the contest outcomes.

2.3.2. Data splitting

The data from the 11 datasets (Table 1) were split into three sets, i.e., training, validation, and testing. The three sets are composed of data from Datasets 1–9, 3–5, and 3–11, respectively. The data in the three sets do not overlap.

The training and validation sets were released to contest participants during the contest to support their method development and refinement. Datasets 10 and 11 were designated as the test set, remaining unseen during the preparation stage, and only made available to participants at the testing stage. The testing set includes not only similar forest types as in the training set, but also an unseen forest type, i.e., tropical open woodland in northern Australia. In the contest, only results from test sets were evaluated to assess the method applicability and transferability.

2.3.3. Evaluation

The evaluation metrics utilized in the instance segmentation include Precision (P), Recall (R), and Average Precision (AP), as shown in (1)–(3).

$$P = \frac{TP}{TP + FP} \times 100\% \tag{1}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{2}$$

$$AP = \int_0^1 PdR \tag{3}$$

where TP, FP, and FN indicate True Positive, False Positive, and False Negative, respectively.

TPs are determined based on the confidence score and Intersection over Union (IoU), which measures the ratio of the intersection to the union of the positive prediction and the reference ground truth as shown in Fig. 2(a). A prediction is considered a TP if its IoU and confidence exceed specified thresholds and the confidence score is the highest if there are redundant predictions to a ground truth ITC mask. As the confidence threshold increases, Precision increases while Recall decreases. The AP represents the integration of the Precision-Recall (P-R) curve over different confidence thresholds.

As shown in Fig. 2(b), for the IoU range between 0.5 and 0.6, an ITC inference is accepted as a TP despite significant crown shape discrepancies with the crown shape of the ground truth. When the IoU threshold is between 0.7 and 0.8, an ITC inference is considered as TP when the inferred crown area aligns closely with the ground truth, despite certain differences in shape or location. When the IoU between 0.9 and 1.0, an ITC inference is accepted as a TP when it nearly perfectly matches the ground truth crown in location, dimension, and shape.

The AP50 was taken as the indicator for the evaluation in the contest. For research purposes, AP75 was also included in the evaluation in this

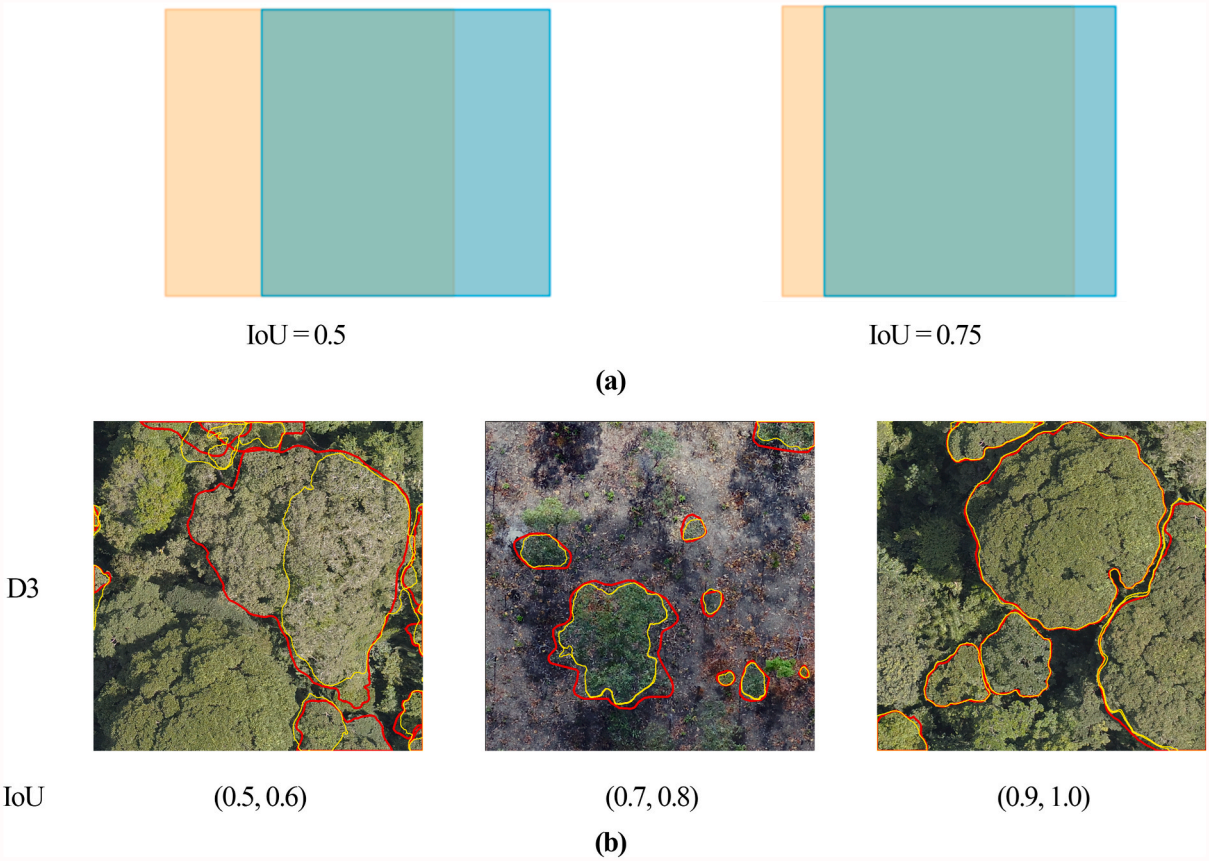

Fig. 2. Examples of different IoU thresholds. (a) The theoretical relationship between the prediction (blue square) and ground truth (orange square), given the IoU value at 0.5 (left) and 0.75 (right). (b) The ITC inferences (yellow polygons) and their corresponding ground truth (red polygons) when the IoU range is set to (0.5, 0.6), (0.7, 0.8), (0.9, 1.0), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

study, to comprehensively investigate the SOTA of the ITD methods and the challenges encountered.

The performance of the crown-size estimation from ITD was evaluated by the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Root Mean Square Error Percentage (RMSE%), and Coefficient of Determination (R^2), as shown in (4)–(7).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - \widehat{a_i}| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \widehat{a_i})^2} \tag{5}$$

$$RMSE\% = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_i - \widehat{a_i}}{a_i}\right)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(a_i - \widehat{a_i})^2}{\sum_{i=1}^{n}(a_i - \overline{a_i})^2} \tag{7}$$

where a_i indicates the tree-level crown area estimates, $\widehat{a_i}$, and $\overline{a_i}$ represent the ground truth and the average value of the ground truths, and n represents the quantity of crowns.

This study benchmarks the ITD methods through "data-based accuracy" based on the annotated reference according to careful human interpretation from images. The gap between data- and reality-based accuracy based on the ground truth in reality is discussed in Section 5.4.2.

2.3.4. The platform

An online platform was established for the contest to manage data distribution, result submission, information release, and feedback. The platform was accessible to all participants during the preparation and testing stages.

During the preparation stage, when a participant submitted a new ITD result, the submission was automatically evaluated, and the ranking was updated in real-time and made available to all participants on the leaderboard. The participants were encouraged to improve their methods based on the feedback and rankings at this stage. In the testing stage, participants submitted their results on the test set. The final evaluation and ranking were based solely on the results from the testing set.

3. Methods

This study investigates four types of methodology for image-based ITD, i.e., the standard DL instance segmentation methods, a general-purpose DL large model, an ML method, and the top six methods in the contest.

This section provides an overview of the top-ranked ITD methods in the contest. Among them, five are newly developed, and the sixth-ranked was the standard Mask DINO with Swin-L backbone. The focus here is on key methodological design elements, such as network architecture, loss functions, and techniques, rather than reviewing each method in exhaustive detail.

Standard DL methods, i.e., Cascade Mask R-CNN, HTC, and Mask DINO, and the large model SAM, are not introduced in detail here because they were applied without modification. More information about these methods can be found in original publications. The ML ITD method is briefly introduced, as no standard ML model exists for image-based ITD.

3.1. Network architectures

The networks simultaneously generate both bounding boxes and masks for ITCs. The typical architecture consists of a series of elements, including a backbone, neck, initial prediction module, and heads. The backbone extracts feature maps from the input images. The optional neck fuses multi-level feature maps by constructing a feature pyramid. The initial prediction module generates bounding box proposals or instance queries. The head produces the results based on the initial predictions and corresponding features. Networks can be categorized into proposal- and query-based methods, depending on the type of initial prediction.

3.1.1. Overview of the individual tree crown segmentation networks

Table 2 provides the detailed information for the top six ITD networks introduced in the contest, highlighting the different modules used at each step across the networks.

According to the baselines, three methods are proposal-based (i.e., Method 1, 4, and 6) and the other three methods are query-based (i.e., Method 2, 3, and 5). Among these methods, four utilized the Transformer and two utilized the ConvNet backbone.

The networks modify the original architecture through various modules. In Method 1, the backbone consists of two identical Swin-L models (Liu et al., 2021) connected by a lightweight Feature Pyramid Network (FPN). In Methods 2 and 3, the Composite Backbone Network (CB-Swin-L) (Liang et al., 2022a) is used as the backbone in the HTC to replace the original ResNeXt-101. Method 2 employs the MaskIoU head from (Huang et al., 2019) for mask generation. Method 4 uses an Efficient Hybrid Encoder (EHE) (Zhao et al., 2024) for high-accuracy and efficient detection, with the DETR with Collaborative Hybrid Assignments Training (Co-DETR) (Zong et al., 2023) and a dynamic convolution-based mask head (Chen et al., 2020). Method 5 replaces the Cascade Mask R-CNN backbone with ConvNeXt (Liu et al., 2022). Method 6 follows the standard Mask DINO as described in its documentation.

3.1.2. Baseline

With the success of the region proposal network (RPN) from Faster R-CNN (Ren et al., 2017), multiple proposal-based instance segmentation networks were gradually proposed, e.g., Mask R-CNN (He et al., 2017), Cascade Mask R-CNN, HTC, etc. The proposals generated by RPN are candidate bounding boxes of objects. Dense rectangular proposals are generated as the regression reference based on the pre-defined anchors

Table 2
The architecture of the top six ITD networks. The modifications to the baseline method are in bold.

Method	Baseline	Backbone	Neck	Anchor network	heads
1	Mask DINO	**Swin-L + FPN + Swin-L**	—	ViT encoder + query decoder	Mask DINO detection + mask head
2	HTC	**CB-Swin-L**	FPN	RPN	HTC detection heads + **MaskIoU head**
3	HTC	**CB-Swin-L**	FPN	RPN	HTC detection + mask heads
4	Mask DINO	**ConvNeXt-v2-L**	FPN	**EHE + Co-DETR**	DINO detection head + **dynamic conv mask head**
5	Cascade Mask R-CNN	**ConvNeXt**	FPN	RPN	cascade conv detection + mask heads
6	Mask DINO	Swin-L	—	ViT encoder + query decoder	Mask DINO detection + mask head

'—' represents without this module.

and feature maps from the backbone. The proposals help to preliminarily locate the object area.

The first query-based network was proposed by an end-to-end object detection with Transformer (DETR) (Carion et al., 2020). Unlike proposal-based networks, it does not rely on predefined anchors or prior knowledge, e.g., the sizes of the detection targets. The learned queries are embeddings generated by the transformer encoder and decoder in the query-based network. The encoder embeds the patches of the feature map. The decoder decodes a fixed-size set of learned object queries in parallel based on the embeddings from the encoder and the initial queries. The learned queries are input to the heads for generating bounding boxes and masks.

3.1.3. Backbone

Transformer-based backbones are popular, e.g., Swin-L and CB-Swin-L. To generalize the Transformer to vision domain, Vision Transformer (ViT) (Dosovitskiy et al., 2021) bridges the gap between language and image by splitting the image into a sequence of patches and considering each patch as an embedding token. It supersedes convolutional network (ConvNet) in image classification. However, its performance is limited in dense vision tasks, e.g., semantic segmentation, detection, and instance segmentation, due to the low-resolution feature map and quadratic complexity with respect to the input data size. Swin Transformer (Swin) proposes a hierarchical structure and a shifted window to generate multi-scale feature maps from fine to coarse resolution by merging nearby patches and cross-patch attentions. Swin takes advantages of Transformer and ConvNet and becomes a backbone for a wide range vision task. It has three versions with different feature scales, in which the Swin-L contains the largest feature dimension.

CBNet proposes the composite connection structure to integrate multiple existing pre-trained backbones to release the burden of pre-training and improve generalization. The CB-Swin-L integrates two Swin-L by the composition connection introduced in CBNet v2 with high-level composition (DHLC) inspired by Dense Convolutional Network (DenseNet) (Huang et al., 2017), auxiliary supervision, and pruning strategies. In DHLC, all features from the current and higher-level stages in the previous backbone are added to the feature in the lower-level stage of the latter backbone.

ConvNeXt improves the ConvNet structure with reference to Swin architecture based on ResNet50. Five primary designs in ConvNeXt include changing stage compute ratio, “Patchify”, ResNeXt-ify, inverted bottleneck, and large kernel size. The changing stage compute ratio modifies the ratio of the number of the stacked blocks into 3:3:9:3. The “Patchify” replaces the first convolution (conv) layer from a 7×7 kernel with stride 2 to a non-overlapping 4×4 conv with stride 4. The ResNeXt-ify utilizes the depth wise conv following the idea of ResNeXt. The inverted bottleneck structure expands the dimension of hidden layers of the block to four times wider than the input. Besides, it takes the 7x7 depth-wise conv as the first layer in each block.

ConvNeXt v2 (Woo et al., 2023) aims to take advantages of the self-supervision learning of masked autoencoders (MAE) (He et al., 2022), and to improve the ConvNeXt v1. The MAE effectively pre-trains the data-hungry encoder of vision Transformer for better generalization.

It is composed of an encoder and a decoder that takes the masked images as input and reconstructs the input images, respectively. After the pre-training, the encoder has a strong generalization as a backbone for multiple down-stream tasks, i.e., classification, detection, instance segmentation.

However, a naïve use of the MAE in ConvNeXt undermines the performance due to the incompatibility of the masked patches of MAE and the dense sliding widows of ConvNets. The fully convolutional masked autoencoder (FCMAE) for MAE pre-trains the encoder in ConvNeXt by applying the sparse conv on visible patches, and the sparse conv is converted back to standard conv during the training and inference. The decoder for MAE pre-training is a lightweight and plain ConvNeXt block to reconstruct the target image. Additionally, a global response normalization (GRN) is introduced to solve the feature collapse caused by dimension expansion in ConvNeXt block by applying global feature aggregation, feature normalization, and feature calibration to increase the contrast and selectivity of channels.

3.1.4. Neck

The neck recovers the feature map hierarchically and symmetrically with the backbone, e.g., through a top-down pathway. The feature pyramid network (FPN) is adopted as the neck in some networks to construct a feature pyramid and fuse multi-scale features. The FPN is originally a detection network (Lin et al., 2017). The top-down feature pyramid, i.e., neck, is built based on the feature hierarchy from backbone with lateral skip-connection. The lateral skip-connections combine low-level features in backbone with fine texture information and high-level features in the neck, which help to discriminate targets. The hierarchical structure contributes to detect multi-scale objects, while it may at the same time introduce a redundant-detection problem over the same target.

3.1.5. Initial prediction module

The instance segmentation networks are categorized in the proposal- and query-based networks. In networks, the results are generated based on the box proposals or learned queries. The initial prediction module aims to generate initial prediction, i.e., proposal or query, as the reference to generate the bounding box and mask.

In the proposal-based network, region proposal network (RPN) is popular as the initial prediction module, introduced in Faster R-CNN (Ren et al., 2017). The RPN generates rectangular object proposals based on feature maps and a set of pre-defined multi-scale anchors, in a sliding window. RPN generates a confidence for each anchor that indicates the probability of enclosing an object, e.g., 1 if the IoU between an anchor and an object is close to 1 and 0 if the IoU is close to 0. These proposals are filtered by a confidence threshold and non-maximum suppression (NMS) to remove redundant anchors. The remaining proposals are input to the head for bounding boxes refinement and mask generation.

In query-based networks, DETR (Carion et al., 2020) is popular. It achieves detection based on a set of learned queries with a predefined number of queries. The initial prediction module of the query-based network is composed by a Transformer encoder and decoder. The encoder is the same as that of ViT, which embeds the feature map patch sequences. The decoder transforms N queries that initially set to 0 to positional encodings referred to as object queries. The object queries are then input to the heads, i.e., feed-forward networks (FFN), as the initial predictions to generate the detection results. However, the bipartite graph matching between queries and ground truths is instable that leads to slow convergence in training.

DN-DETR stabilizes training and accelerates convergence by feeding noised ground truth (GT) to the decoder (Li et al., 2022). The noise levels on center shifting and box scaling are controlled by two hyper-parameters. The Deformable DETR introduces a deformable attention module inspired by deformable convolution (Dai et al., 2017; Zhu et al., 2018) to improve the query learning, where only top-k nearest embeddings around a reference are assigned as keys for each query in the Transformer attention models.

The DETR with improved denoising anchor boxes (DINO) is another improved version of DETR that aims to improve the denoising training, query initialization, and box prediction (Zhang et al., 2022). The contrastive denoising training enhances the model ability to distinguish slight differences between queries and avoid duplicate predictions by setting two hyper-parameters for positive and negative noised queries generation. The mixed query selection improves the deformable attention module in Deformable DETR by only enhancing the position embedding with top-k nearest embeddings and keeping the content queries learnable as in DETR, enabling learning more features from the encoder. In addition, look forward twice improves the iterative box refinement by updating the parameters of each layer in the head based

on the loss of itself and previous layer.

Collaborative hybrid assignment training scheme (Co-DETR) (Zong et al., 2023) is introduced to enrich the supervision of encoder and attention learning in decoder. The collaborative hybrid assignment training manages to enrich the localization supervision of the encoder by incorporating versatile auxiliary heads and one-to-many label assignment manner to exploit multiple object queries. Multiple auxiliary heads are compared, where the combination of adaptive training sample selection (ATSS) + Faster R-CNN performs the best. The auxiliary heads are applied only during training to enrich the supervision on the encoder's output to support the training convergence, and are discarded in inference. This training scheme can be generalized to multiple query-based detection network, e.g., Deformable-DETR++, DINO, etc. Additionally, sufficient customized positive queries are generated in each auxiliary head to enhance the cross-attention learning in decoder.

Real-time detection transformer (RT-DETR) (Zhao et al., 2024) introduced the efficient hybrid encoder (EHE) to replace the vanilla Transformer encoder. EHE is composed by the attention-based intra-scale feature interaction (AIFI) and the CNN-based cross-scale feature fusion (CCFF). The AIFI applies the self-attention operation to the high-level feature map, and the CCFF fuses the multi-level feature maps from two adjacent layers. RT-DETR processes multi-scale features more efficiently.

3.1.6. Heads

The network head generates both ITC bounding boxes and masks. Hence, the instance segmentation network includes detection and segmentation heads. Three classic instance segmentation baseline networks are adopted in the top-ranked methods, i.e., Cascade Mask R-CNN, hybrid task cascade (HTC) (Chen et al., 2019), and Mask DINO (Li et al., 2023). The cascade structure is initially introduced in Cascade R-CNN (Cai and Vasconcelos, 2021) to achieve progressive refinement. Both Cascade Mask R-CNN and HTC employ the cascade structure.

In Cascade R-CNN, a sequence of detection heads are combined. The bounding boxes output from the previous head are the input to the next head as box proposals. Thus, the proposals from the RPN are progressively refined by continuous heads. The Cascade Mask R-CNN is the naïve combination of the Cascade R-CNN and the conv mask head from Mask R-CNN, which generate bounding boxes and masks in parallel. The detection head is a conv detection head in Faster R-CNN.

In Cascade Mask R-CNN, there is no interaction between mask heads. In HTC, the intermediate features from the previous mask head are the input to the next mask head for feature fusion, achieved by an element-wise sum, and thus to refine masks progressively. In addition, a semantic segmentation branch predicts a pixel-wise semantic map for whole image, and pass the semantic feature to each mask head.

Mask DINO is a query-based instance segmentation network which extends the DINO by adding a mask branch. The detection head is the feed forward network (FFN) in DINO. The pixel-level embedding for mask generation is constructed based on the feature maps recovered from the encoder and backbone, inspired by Masked-attention Mask Transformer (Mask2Former) (Cheng et al., 2022). The resolution is recovered by up-sampling, and the integration of the feature maps is achieved from the encoder and backbone by pixel-wise sum. Each content query computes the dot product with the pixel-wise embedding to obtain a mask embedding.

3.2. Loss function

Loss function is one of the most influential components for network training. The network parameters are optimized based on the gradient descent of loss function. The loss function of the instance segmentation network is composed of detection and segmentation loss.

In HTC, the loss function integrates the detection loss $\mathscr{L}^t_{det}$, mask loss $\mathscr{L}^t_{mask}$, semantic segmentation loss $\mathscr{L}_{seg}$, as shown in (8)–(11). The $\mathscr{L}^t_{det}$ integrates the classification loss $\mathscr{L}_{cls}$ and box loss $\mathscr{L}_{box\_L_1}$. Given T hybrid task heads, at each stage t, the detection head predicts the classification score c_t, offsets between proposals and GT bounding boxes r_t, and the mask head predicts proposal-wise mask m_t. The semantic segmentation head predicts a full-image pixel-wise semantic map s.

$$\mathscr{L}_{HTC} = \sum_{t=1}^{T} \alpha_t \left(\mathscr{L}^t_{det} + \mathscr{L}^t_{mask}\right) + \beta \mathscr{L}_{seg} \tag{8}$$

$$\mathscr{L}^t_{det}(c_t, r_t, \hat{c}_t, \hat{r}_t) = \mathscr{L}_{cls}(c_t, \hat{c}_t) + \mathscr{L}_{box\_L_1}(r_t, \hat{r}_t) \tag{9}$$

$$\mathscr{L}^t_{mask}(m_t, \hat{m}_t) = BCE(m_t, \hat{m}_t) \tag{10}$$

$$\mathscr{L}_{seg} = CE(s, \hat{s}) \tag{11}$$

where $\hat{c}_t$, $\hat{r}$, $\hat{m}_t$, $\hat{s}$ represent the GT, c_t, r, m_t, s represent the predictions, $\mathscr{L}_{ce}(c_t, \hat{c}_t)$ is the classification cross-entropy (CE), $\mathscr{L}_{box\_L_1}(r_t, \hat{r}_t)$ is the L1 loss function, $\mathscr{L}^t_{mask}$ is the binary cross entropy (BCE), $\mathscr{L}_{seg}$ is the segmentation CE. The coefficients are hyper-parameters, with $\alpha = [1, 0.5, 0.25]$, T = 3, and $\beta = 1$.

In Mask DINO, the loss function integrates detection loss $\mathscr{L}_{det}$ and mask loss $\mathscr{L}_{mask}$ to evaluate the gaps between the predictions and GT, as shown in (12–14). The $\mathscr{L}_{det}$ integrates the classification loss $\mathscr{L}_{cls}$ and two box loss $\mathscr{L}_{box\_L_1}$ and $\mathscr{L}_{box\_giou}$.

$$\mathscr{L}_{MaskDINO} = \mathscr{L}_{det} + \mathscr{L}_{mask} \tag{12}$$

$$\mathscr{L}_{det} = \lambda_{cls}\mathscr{L}_{cls} + \lambda_{box\_L_1}\mathscr{L}_{box\_L_1} + \lambda_{box\_giou}\mathscr{L}_{box\_giou} \tag{13}$$

$$\mathscr{L}_{mask} = \lambda_{bce}\mathscr{L}_{bce} + \lambda_{dice}\mathscr{L}_{dice} \tag{14}$$

where the classification loss $\mathscr{L}_{cls}$ is the focal loss (Ross and Dollár, 2017), the detection loss integrates the L1 loss function $\mathscr{L}_{box\_L_1}$ and GIoU loss $\mathscr{L}_{box\_giou}$ (Rezatofighi et al., 2019). The mask loss combines BCE $\mathscr{L}_{bce}$ and dice loss $\mathscr{L}_{dice}$. The coefficients are set to $\lambda_{cls} = 4$, $\lambda_{box\_L_1} = 5$, $\lambda_{box\_giou} = 2$, $\lambda_{bce} = 5$, and $\lambda_{dice} = 5$.

In the contest, Method 2 and 3 use the original loss function of HTC, Method 5 uses the loss function of Cascade Mask R-CNN, Method 6 uses the original loss function of Mask DINO, Method 1 replaces the dice loss to focal tversky loss (Abraham and Khan, 2019), and Method 4 adds a varifocal loss (Zhang et al., 2021) in detection loss.

3.3. Technical enhancements

Several technical enhancements are applied to improve ITD, i.e., data augmentation, training strategies, and inference enhancements. Table 3 summarizes the technical enhancements used in the networks.

Data augmentation is a set of operations to increase the quantity of the training set by image transformations, e.g., shape transformation and value changing. They help to improve the robustness and generalization of the instance segmentation method. The shape transformation

Table 3
The application of technical enhancements in the ITD networks.

Method	Data augmentation	Training strategy				Test time augmentation
		Optimizer	GT	AMP	EMA	
1	√	AdamW	√	√	—	—
2	√	AdamW	—	—	√	√
3	√	AdamW	—	—	—	—
4	√	AdamW	√	—	√	√
5	√	SGD	—	√	—	√
6	√	AdamW	—	—	—	—

'—' and '√' represent without and with these operations, respectively; AdamW, SGD, GT, AMP, and EMA represents Adam (Adaptive Moment Estimation) with decoupled weighted decay, stochastic gradient descent, gradient clipping, automatic mixed precision, and exponential mixed precision, respectively.

changes the images and their corresponding labels by multiple operations, e.g., resizing, cropping, flipping, and rotation. The value changing adjusts the digital numbers of the images by, e.g., adding noise, blurring, changing brightness and contrast, and transforming hue and saturation. All of the top-ranked methods applied the data augmentation during data pre-processing.

In network training, the optimizer is important for the gradient descent. The stochastic gradient descent (SGD) and AdamW were utilized by the method 5 and the rest, respectively. Several ITD methods utilized multiple solutions to improve the stability and efficiency in training, e.g., gradient clipping, automatic mixed precision, exponential moving average.

The test time augmentation (TTA) (Moshkov et al., 2020) produces the average outputs based on the augmented testing set. Multiple top-ranked methods applied this augmentation to enhance the inference. TTA has four components, i.e., augmentation, prediction, dis-augmentation, and merging. The augmentation process applies multiple operations, e.g., flipping, rotation, and resize, on the testing images. The inference is performed on both original and augmented images. The de-augmentation transforms the masks in augmented images back to the original images. And, the masks of the same object are merged according to IoU threshold. TTA is a risky strategy that leads to accuracy improvement but may reduce several correct predictions (Shanmugam et al., 2020).

3.4. Ml-based ITD method

The watershed-based method manages to achieve ITD based on the geometric and radiometric characteristics that the ITCs contain high radiation intensity due to their high, uppermost sunlit portions (Wang et al., 2004; Wang, 2010; Jing et al., 2012; Chemura et al., 2015). This study applies marker-controlled watershed segmentation with morphological pre-processing in Matlab (The MathWorks Inc, Massachusetts, United States) to explore the performance of the ML-based ITD method, as there is no standard pipeline. The marker-controlled watershed segmentation delineates objects by growing the "catchment basins" as object areas and "watershed ridge lines" as boundaries based on the foreground and background markers. However, directly applying marker-controlled watershed segmentation typically leads to severe over-segmentation. The morphological operations, i.e., opening-by-reconstruction and closing-by-reconstruction, are applied in pre-processing to reduce over-segmentation by creating flat maxima inside each object.

4. Results

The results are presented using both visual verification and evaluation metrics. The evaluation metrics of the methods in the contest were calculated by the operating dockers submitted by participants, in which the ITC inference results were submitted within.

4.1. Illustrations of individual-tree-crowns masks and the determination of tree-positive instances

Fig. 3 presents examples of ITC inference masks generated by the top 6 ranking methods in the contest for each study site. The visualized ITD outcomes from each method across datasets offer an intuitive understanding of forest types, stand conditions, and the accuracy of ITC masks with reference to the ground truth. All top-ranked methods have redundant ITC predictions, i.e., multiple predictions for a single target.

Fig. 4 illustrates the differences between TPs corresponding to different IoU thresholds, namely $IoU \geq 0.5$ and $IoU \geq 0.75$, illustrated by the results from methods ranking 1st and 6th in the contest. The results show a significant decrease in TPs (fewer yellow polygons) for both methods when using a higher IoU threshold, particularly in more challenging datasets such as D4, D8, and D11.

As shown in Fig. 4, the TPs corresponding to a higher IoU threshold ($IoU \geq 0.75$) correspond the ground truth more accurately, in terms of mask position, size, and morphological fidelity in comparison with the TPs corresponding to a lower IoU threshold ($IoU \geq 0.50$). Meanwhile, the recall of TPs decreases sharply with a higher IoU threshold.

This result suggests that the selection of the accuracy metrics should be application dependent. In applications where the counts or locations are the expected outcomes, AP50 probably gives a good approximation of the crown segmentation. However, the AP75 and higher values, instead of the popular AP50, should be utilized as an accuracy indicator in applications where the actual boundary properties are important, e.g., to study the site structure or to estimate tree attributes according to the crown size. This result also suggests that research and application are domain dependent. The well-accepted golden rules or knowledge in one area may not suit the requirements in another area.

4.2. The evaluations of deep-learning method performances

The evaluations are divided into two groups. The first group includes the SOTA ITD methods that were not involved in the contest. These methods were implemented using the dataset from the contest. The second group includes the top 6 ITD methods of the contest. All methods were evaluated using identical assessment approaches and standards, so that the performances of all methods can be compared in and outside of the contest.

4.2.1. The existing methods

The performance of several SOTA instance segmentation networks in ITD is verified. The AP50 and AP75 of the inference results are shown in Tables 4 and 5, respectively. The mean AP50 and AP75 are approximately 40% and 19%, respectively, except for SAM that is clearly lower than others.

The performances of HTC, Mask DINO, and Cascade Mask R-CNN are similar. Cascade Mask R-CNN and MaskDINO gain most of the highest score in AP50 and AP75, respectively. The transferability of Cascade Mask R-CNN is the best, as it outperforms others in both testing datasets that are not involved in the training and validation, i.e., Datasets 10 and 11.

The SAM performed the worst in the benchmark. The generalization of the networks trained by domain datasets significantly outperforms the zero-shot inference of SAM.

4.2.2. The methods from the contest

Fig. 5 illustrates the average AP50 and AP75 across all datasets for the top 20 teams of the contest, providing insight into the overall ITD performance. With AP50 as the indicator, five teams achieved an average accuracy above 50% when averaged across all datasets. However, when AP75 was applied, accuracy dropped significantly, with only 7 top-ranked teams exceeding 25% and none surpassing 30%. This indicates that achieving AP75-level alignment between ground truth and ITC inferences remains a significant challenge for all methods.

Fig. 6 illustrates the AP50 scores of the top 20 teams for each dataset, showing significant variation across sites. In Dataset 6, the top 6 ITD methods achieved AP50 values of approximately 75%, the highest among all datasets, while in Dataset 4, the values are around 20% even with the best performing methods, the lowest recorded.

Although all methods performed variably across different datasets, strong performance patterns can be observed, particularly among the top 6 methods (Methods 1–6 in Figs. 5 and 6). These methods consistently performed well in certain datasets (i.e., Dataset 3, 5, 6, 7, 9, and 10) and struggled in others (i.e., Dataset 4 and 8). A notable exception is Dataset 11 (Savanna woodland), where Methods 3, 4, and 6 showed significantly lower performance, indicating a weakness of those methods in handling this specific stand type. In contrast, methods ranked 10–20th exhibited inconsistent performances across datasets, suggesting greater sensitivity to site-specific characteristics.

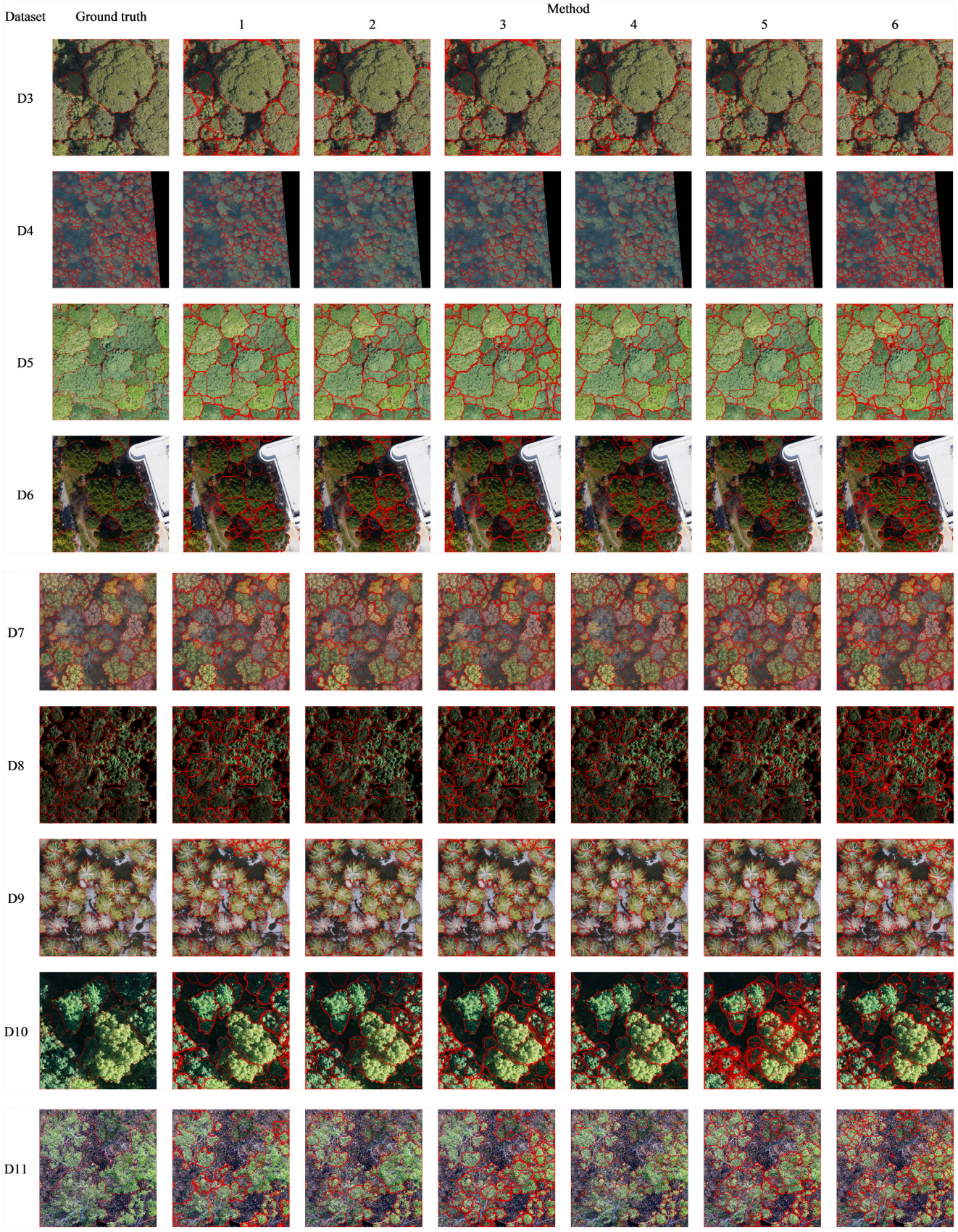

Fig. 3. Examples of the individual-tree-crown (ITC) masks. The first column illustrates the ground truth. The column 2–7 represents the inference masks from the top 6 methods. D3–11 represent the Dataset 3 to 11 in the testing set. All top-ranked methods have redundant ITC predictions.

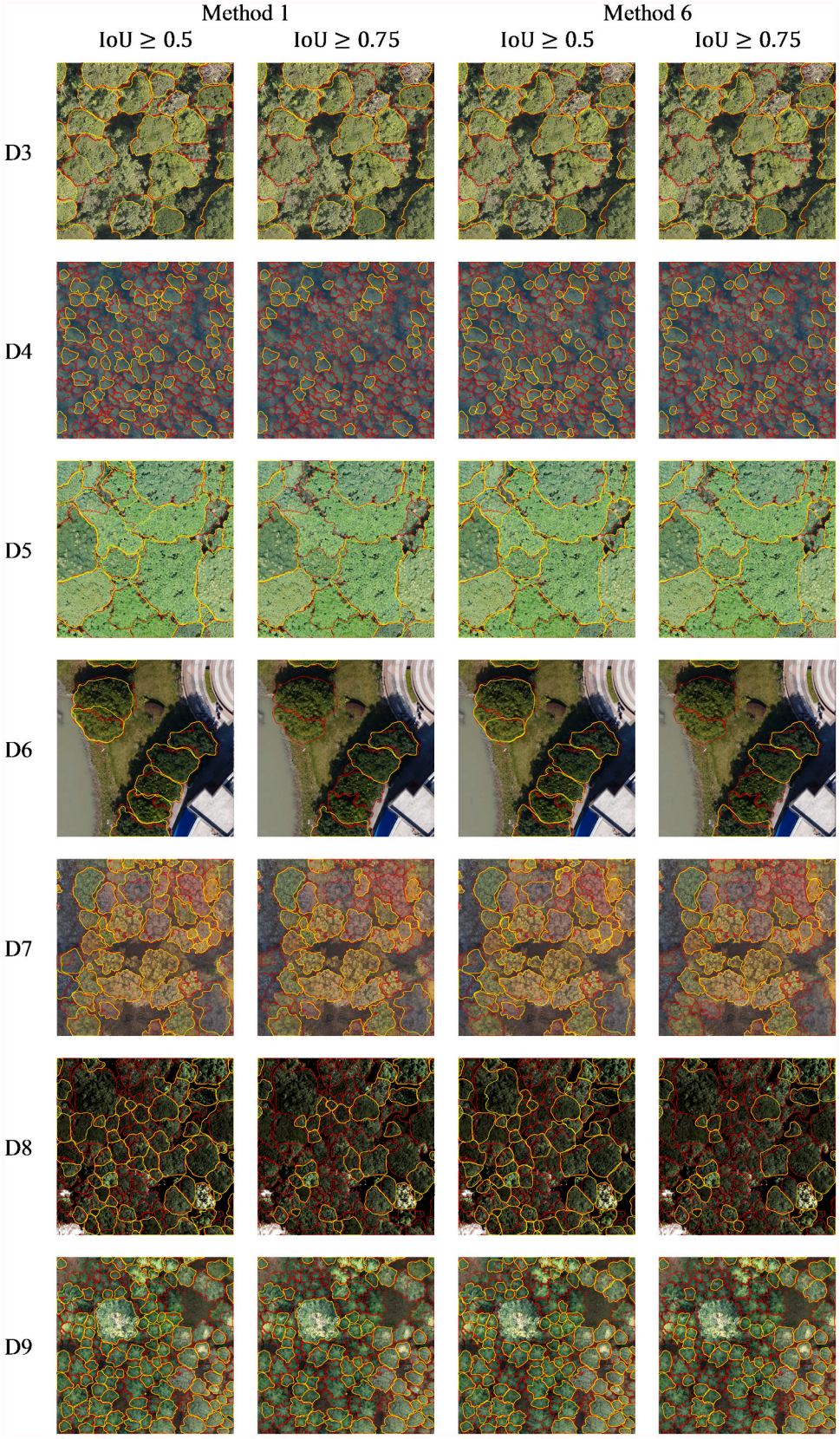

Fig. 4. The TP ITD results according to 0.5 and 0.75 IoU threshold. The red and yellow polygons represent the ground truth and TP prediction, respectively. The Methods 1 and 6 are the 1st and 6th ranked methods, respectively, in the contest. D represents the Dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

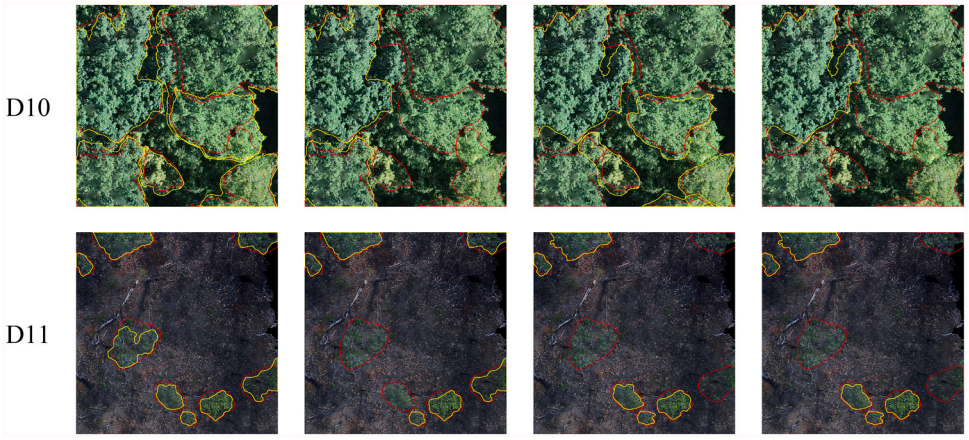

Fig. 4. (*continued*).

Table 4
The AP50 of the state-of-the-art instance segmentation methods. The highest AP50 of each dataset is highlighted in bold.

Method	mean AP50	Dataset								
		3	4	5	6	7	8	9	10	11
SAM	8.47	8.80	13.51	2.99	3.22	22.76	10.73	12.33	0.70	1.19
HTC	39.05	60.10	**18.68**	49.08	68.15	60.91	34.03	46.82	11.57	2.09
MaskDINO	39.08	56.35	18.64	49.75	**69.03**	59.81	33.10	**50.80**	10.91	3.31
Cascade Mask R-CNN	**40.66**	**60.98**	18.61	**49.90**	64.03	**64.93**	**35.84**	46.88	**18.69**	**6.04**

Table 5
The AP75 of the state-of-the-art instance segmentation methods in benchmark. The highest AP75 of each dataset is highlighted in bold.

Method	Mean AP75	Dataset								
		3	4	5	6	7	8	9	10	11
SAM	3.83	4.81	6.20	1.56	2.11	12.22	4.22	2.93	0.22	0.20
HTC	19.57	34.83	8.81	21.81	45.54	29.79	14.21	18.09	2.50	0.59
MaskDINO	**19.83**	32.43	**9.35**	**22.56**	**46.24**	29.45	**14.44**	**20.85**	2.56	0.57
CascadeMask R-CNN	19.66	**36.09**	7.87	21.38	40.80	**30.04**	14.31	19.01	**5.47**	**1.99**

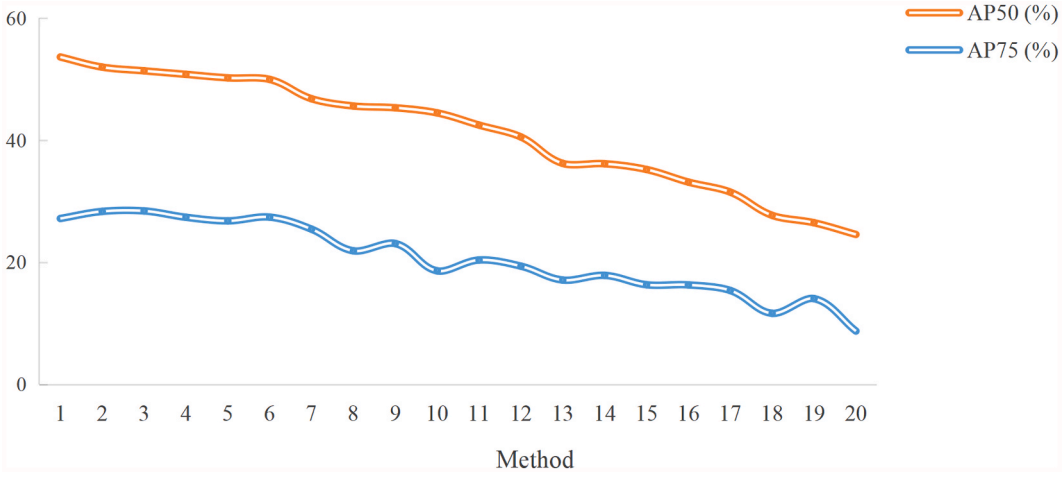

Fig. 5. Results of the top 20 methods in the testing phase.

Tables 6 and 7 list the AP50 and AP75 scores of the top 6 methods, respectively, based on the final ITC inferences results submitted at the evaluation stage of the contest.

Table 6 indicates that, for AP50, Methods 2 and 3 achieved the best performance on three datasets, where Method 2 on sites 5, 8, and 9 and Method 3 on sites 4, 6, and 10. For AP75 (Table 7), Method 3 maintained its top performance on the same three datasets, while Method 2 led on sites 3 and 5. However, Method 3 performed poorly on Dataset 11, significantly lowering its overall AP50 average. In contrast, Method 1 excelled in Dataset 11 and maintained strong performance across other sites, resulting in the highest average AP50 score. Nevertheless, despite its weaker performance in Dataset 11, Method 3 still achieved the

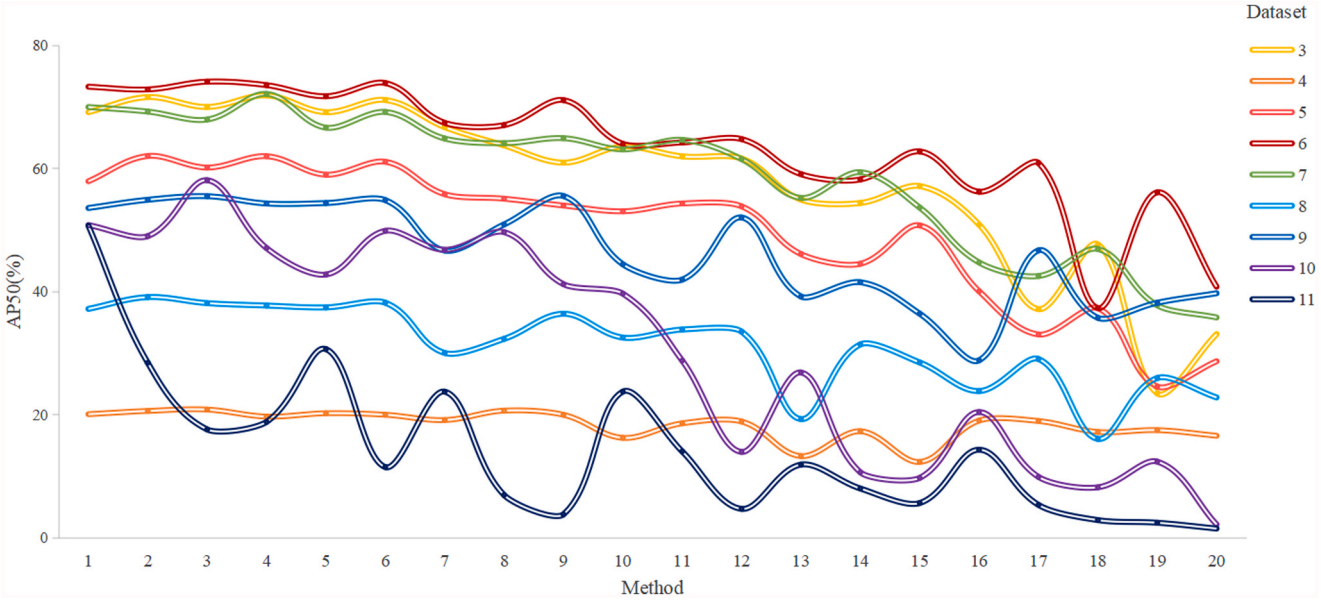

Fig. 6. The AP50 scores of the top 20 methods in the contest across different datasets.

Table 6
The AP50 values of the top 6 methods in the testing set. The highest AP50 is highlighted in bold.

Method	mean AP50	Dataset								
		3	4	5	6	7	8	9	10	11
1	**53.66**	69.18	20.09	57.94	73.31	70.01	37.21	53.60	50.83	**50.75**
2	51.99	71.61	20.62	**62.05**	72.87	69.27	**39.12**	**54.93**	49.03	28.45
3	51.64	70.71	**20.85**	60.74	**74.04**	68.79	38.82	55.55	**57.81**	17.49
4	50.82	**71.89**	19.78	62.02	73.58	**72.13**	37.75	54.33	47.10	18.84
5	50.23	69.17	20.25	59.01	71.74	66.65	37.45	54.41	42.72	30.67
6	49.99	71.15	20.01	61.10	73.90	69.22	38.21	54.89	49.90	11.50

Table 7
The AP75 values of the top 6 methods the testing set. The highest AP75 is highlighted in bold.

Method	mean AP75	Dataset								
		3	4	5	6	7	8	9	10	11
1	27.21	46.28	8.69	30.02	49.96	36.09	15.02	22.30	23.50	**13.01**
2	28.36	**48.64**	10.77	**33.73**	49.98	37.17	18.28	24.20	22.62	9.84
3	**28.45**	46.90	**11.18**	32.75	**51.85**	37.20	17.19	24.50	**28.47**	5.98
4	27.42	46.99	9.32	32.18	49.92	38.46	16.82	**27.16**	18.23	7.72
5	26.82	43.33	10.30	29.22	49.18	33.57	**18.30**	25.88	19.54	12.06
6	27.43	48.16	9.52	32.52	50.81	**38.60**	16.24	25.63	22.01	3.36

highest average AP75 score across all datasets.

The overall performance trend of the top 6 ranking methods is consistently and remarkably better than existing SOTA methods across datasets (Tables 6), despite variations in exact AP scores (Tables 7). All methods performed similarly and tended to perform better on certain datasets and worse on others. The exceptions took place in forest types or conditions that have not been trained before testing, i.e., Datasets 10 and 11, where AP50 scores presented significant variation among methods. For most datasets, however, the performance gap between the best and worst methods is less than 2%. Significant AP50 variances were seen in Datasets 10 and 11 where Method 3 and 1 outperformed the others, respectively, suggesting better transferability of these methods to these specific environments.

4.2.3. The individual-tree-crown measurement

The ITC parameters, e.g., crown size, provide tree-level canopy descriptions. To reveal the SOTA of the ITD methods, the crown-size estimation is evaluated based on the TP.

Tables 8 and 9, as well as Fig. 7, indicate the performance of tree-level crown-area estimation from the Method 1 in the contest under different IoU thresholds, i.e., 0.50 and 0.75. The recall is significantly higher than precision, except for the Dataset 4. For most datasets, the redundant predictions result to poor performance in precision, however, contribute to high recall. Dataset 4 has the highest precision but lowest recall, showing reversed trends with other datasets.

The IoU threshold impacts the performance of the crown-area estimation. The MAE, RMSE, RMSE%, and R^2 show improvement with a stricter IoU threshold. A higher IoU threshold negatively impacts precision and recall values, but significantly improves the reliability of the TPs and the accuracy of the tree-level crown-area estimation.

4.2.4. A comparison of the existing state-of-the-art methods and methods from the contest

Fig. 8 illustrates the performance of the top 6 ranking methods in the

Table 8
The performance of the crown-area estimation based on the TPs from Method 1, where IoU is larger than 0.5. The highest and lowest values are highlighted in bold.

	Dataset								
	3	4	5	6	7	8	9	10	11
Precision (%)	13.24	**45.14**	16.27	7.65	25.20	33.45	31.64	**3.88**	5.99
Recall (%)	88.04	**21.16**	73.59	86.00	78.68	46.25	60.67	**89.09**	75.19
MAE (m^2)	6.36	4.49	2.09	2.03	5.91	**14.23**	**0.89**	6.43	3.95
RMSE (m^2)	12.71	6.88	4.37	4.93	10.78	**23.14**	**1.68**	11.89	7.46
RMSE%	23.34	27.84	**30.06**	24.65	**22.61**	27.50	23.13	28.75	28.07
R^2	**0.96**	**0.81**	0.91	0.91	0.90	0.85	0.94	0.89	0.90

Table 9
The performance of the crown-area estimation based on the TPs from Method 1, where IoU is larger than 0.75. The highest and lowest values are highlighted in bold.

	Dataset								
	3	4	5	6	7	8	9	10	11
Precision (%)	10.04	**27.59**	9.95	5.53	15.82	18.35	18.56	**2.40**	2.53
Recall (%)	**66.81**	**12.93**	45.01	62.20	49.39	25.36	35.58	55.10	31.77
MAE (m^2)	3.70	2.60	1.11	1.11	**3.80**	7.29	**0.60**	3.70	3.36
RMSE (m^2)	6.79	3.76	2.23	2.16	6.12	**11.73**	**1.06**	6.33	5.98
RMSE%	**10.61**	12.99	11.19	10.65	11.32	12.50	10.97	12.25	**15.80**
R^2	**0.99**	**0.95**	0.98	0.98	0.97	0.96	0.98	0.97	0.96

contests as well as the four SOTA methods across testing Datasets.

Three out of four other SOTA methods exhibited a similar performance trend across datasets. The exception was the SAM, whose zero-shot inference struggled to distinguish ITCs from other foreground instances. Its regular grid vertices as seed points constrained ITC localization, and its transferability was very limited concerning forest conditions. These highlight the importance of re-training models with appropriate data when applying them to new use cases like ITD.

The top 6 contest methods showed significant improvements in Datasets 5 (sub-tropical evergreen broadleaf forest) and 10 (temperate mixed forest) in comparison with the SOTA methods, indicating notable progress can be achieved through method design in ITD for broadleaf trees.

4.3. Results based on influencing factors

The ITD performances were influenced by multiple factors, including the data quality, forest type, and method. To better understand their impacts on ITD performance, the results were analyzed in detail based on key factors such as image quality and climate zone. The analyses are based on the ITD results without any additional process, e.g., post-processing that removes redundant predictions and refines the ITD results.

4.3.1. Image resolution

Fig. 9 illustrates the impacts of the image resolution, according to the AP50 values and top 20 methods.

The image-resolution effects are revealed by the agreement among the overall performances in datasets with similar resolutions despite distinct forest types. As shown in Fig. 9, both Dataset 4 and 8 present significant lower performances across all compared methods, which have 10 cm resolution, while Datasets 3, 5, 6, 7, and 9 whose resolutions are higher than 5 cm have clearly better performances in comparison with the Dataset 4 and 8, although the actual performances are affected by both data quality and forest conditions.

Fig. 10 visualizes the ITD results of Datasets 4, 5, and 7, where image resolutions range from 2 to 10 cm, taking results from Method 1 and Method 6 as examples. These illustrate how higher image resolution generally enhances ITD performance for similar canopy structures.

4.3.2. Image quality – sharpness, shading, contrast

Fig. 11 presents the ITD results of three datasets with similar image resolution but varying image quality due to differences in image sharpness, shading effects, and contrast. Despite similar resolution levels and canopy structures, Datasets 4 (10 cm resolution), 8 (10 cm resolution), and 9 (3–7 cm resolution) exhibit significantly different overall performances, with an approximately 20% average accuracy gap across most of the 20 methods in the contest.

Fig. 12 visualizes the ITD results of Datasets 4, 8 and 9, taking results from Method 1 and Method 6 as examples. The datasets are with varied image quality, where Dataset 4 is blurry (blur), Dataset 8 shows intensive shading besides big crowns (dark), and Dataset 9 presents a satisfactory level of sharpness and contrast (high quality). In Datasets 4 and 8, where the image resolutions are the same and the canopy structures are similar, the method performances suggest that the mitigation of shading effects is easier than mitigating image blurriness. Results in Dataset 9 indicate that image sharpness contributes to a better ITD performance.

4.3.3. Forest conditions

Significant variations exist among forest scenes across different climate zones and forest types. Fig. 13 presents the AP50 scores from the top 20 methods across four forest conditions, i.e., tropical natural forest, subtropical natural forest, subtropical urban forest, and temperate natural forest. The resolutions of Datasets 5, 6, 10, and 11 are all at a level of 2–3 cm (2, 3, 2, and 2 cm, respectively), with sufficient image quality to clearly depict ITC boundaries and texture. The results suggest that the current ITD methods perform well for sub-tropical natural and urban forests (Datasets 5 and 6), have difficulties in temperate nature forests, and are greatly challenged by tropical nature forests.

Fig. 14 visualizes the ITD results of four datasets with similar resolution and image quality but from different climate zones, taking results from Method 1 and Method 6 as examples. It is worth noting that the transferability significantly impacts the method performance, besides the overall canopy structure. Most methods except Method 1 have difficulties in the Datasets 10 and 11, because they were excluded during the training stage and only provided at the evaluation stage. More discussions about the method transferability are in section 5.5.

4.4. Results of ML-based method

Fig. 15 visualizes the ITD results of a watershed-based ML ITD method. Results have clear over-, under- and miss-segmentation, as well as confusion.

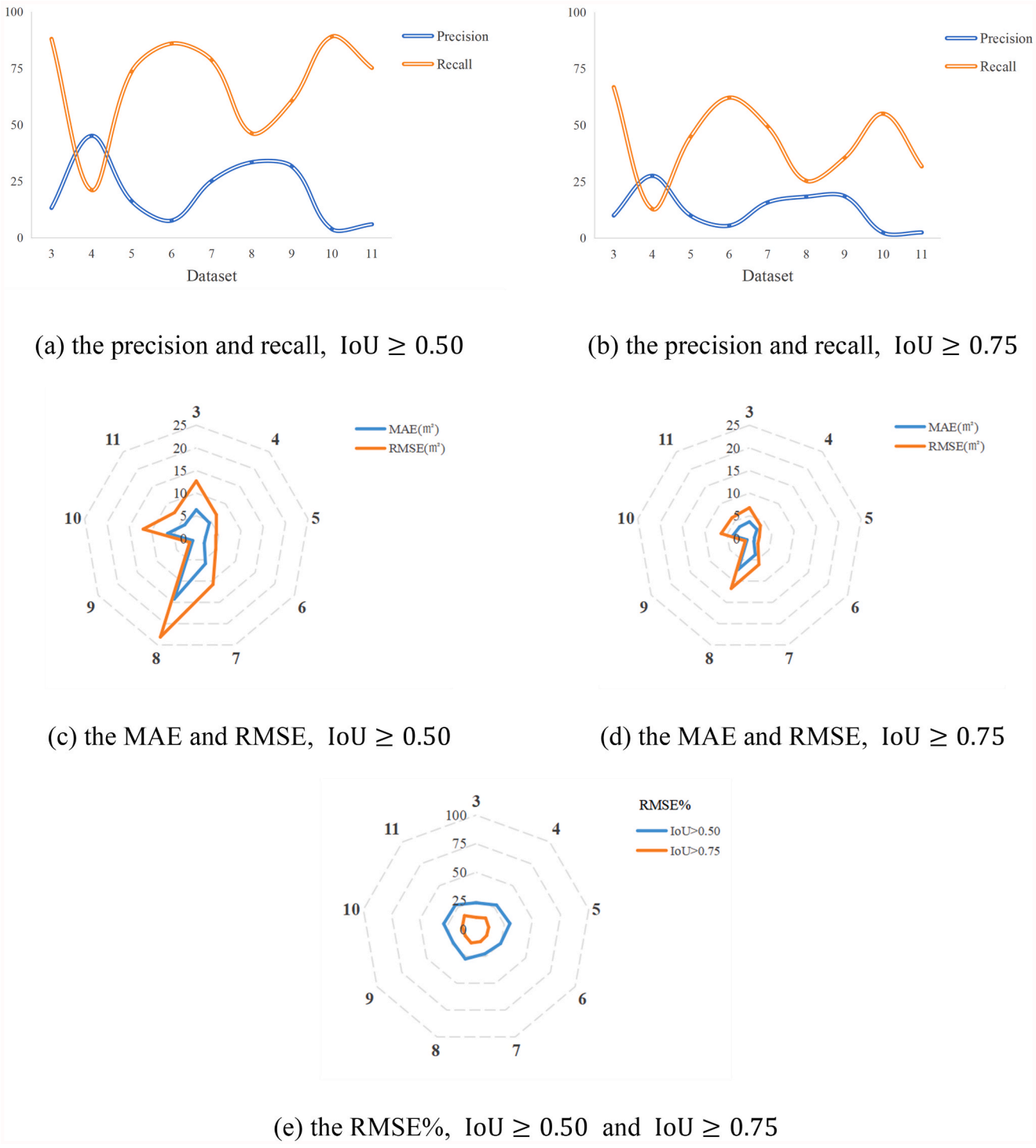

Fig. 7. The performance of the tree-level crown-area estimation based on the TP from the Method 1 in the contest.

The performance relies on the parameters of the methods and texture heterogeneity. Disk size is a parameter that significantly impacts the morphological dilation and erosion operations in pre-processing, which is the radius of the flat disk-shaped morphological structuring element. Size 5 is too small for most datasets causing over-segmentation, which, however, manages to identify most crowns in Dataset 4, as shown in Fig. 15(a). Larger disk size (i.e., 10 and 20) accurately delineates several crowns with apparent texture heterogeneity between background and adjacent crown, as shown in Datasets 3, 4, 5, 7, 9, 10 in Fig. 15(b), and Dataset 3, 4, 5, 7, 9 in Fig. 15(c). However, clear miss-segmentation existed due to crown size variety and fixed structural size. Under-segmentation was shown in Datasets 3, 5, 7, and 9 in Fig. 15(b). Larger disk size tends to delineate the boundaries between crown clumps and background, but fails to separate adjacent crowns with similar texture, as shown in Datasets 3, 6, 8, and 10 in Fig. 15(c). Additionally, segmentation confusion appears in Dataset 6 and 11, where the method fails to distinguish ITCs from other no-disk-like objects in the foreground and segments the background with complex texture into fragments.

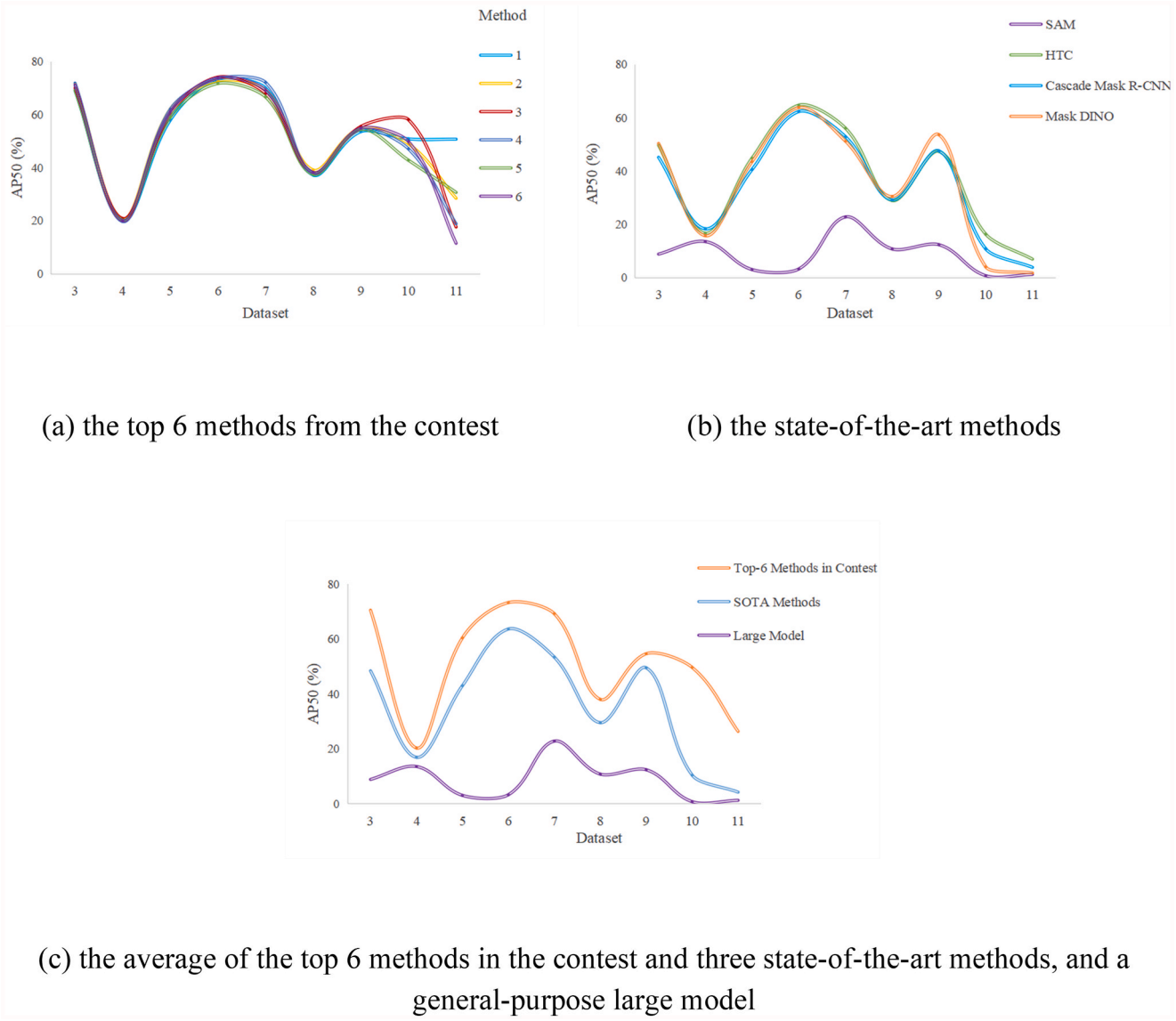

Fig. 8. The AP50 of instance segmentation methods in the testing set.

5. Discussion

This section discusses the ITD results based on the main factors that influence performance. The impacts of evaluation criteria on result credibility, as well as the transferability and generalization of the ITD methods are also introduced.

5.1. Data quality

5.1.1. Image resolution

Image resolution directly affects the visibility and sharpness of object texture and edges. Generally, lower resolution leads to less accurate ITD due to unclear or blurred image textures, as seen in Fig. 9. AP scores were lower in low-resolution datasets, e.g., Datasets 4 and 8 with 10 cm resolution, in comparison with those in high-resolution datasets, e.g., Datasets 3, 5, 6, 7, and 9 with 5 cm resolution or higher.

Moreover, accuracy variances are also observed among datasets with similar image resolutions. In high-resolution datasets (higher than 5 cm), Datasets 3, 6, and 7 have similar accuracy, while Datasets 5 and 9 have relatively lower accuracy. Both Datasets 4 and 8 have 10 cm resolution, yet their AP50 scores are at different levels across all methods as shown in Fig. 9.

These findings suggest that high image resolution does not guarantee higher ITD accuracy, as other factors, such as forest types, also play a crucial role. For instance, Dataset 5, despite having the highest resolution, shows lower AP50 scores in comparison with other datasets with resolution higher than 5 cm, likely due to complex forest conditions such as high tree density and homogeneous canopy textures. Similarly, the lower AP50 scores for Dataset 9 may be attributed to scene diversity, i.e., the images were collected in multiple locations sparsely distributed over southeastern Norway, and the image resolution ranges from 2 cm to 7 cm. Therefore, Dataset 9 presents varied scenes for the same forest type, which challenges the ITD. Nevertheless, when the forest type is similar, datasets with higher resolution generally yielded better ITD accuracy, as observed in Datasets 3, 6, and 7, in comparison with Datasets 5 and 9.

5.1.1.1. Image quality. When image resolutions and forest types are similar, image quality plays a crucial role in the accuracy of ITD, as shown in Fig. 11. Image quality primarily refers to factors like visibility and sharpness, lighting condition, scene clarity, shading, and contrast, which all impact the overall visual quality of the image. For example, Datasets 4 and 8 have similar image resolution, however, the images in Dataset 4 are significantly blurry compared to those in Dataset 8. Such

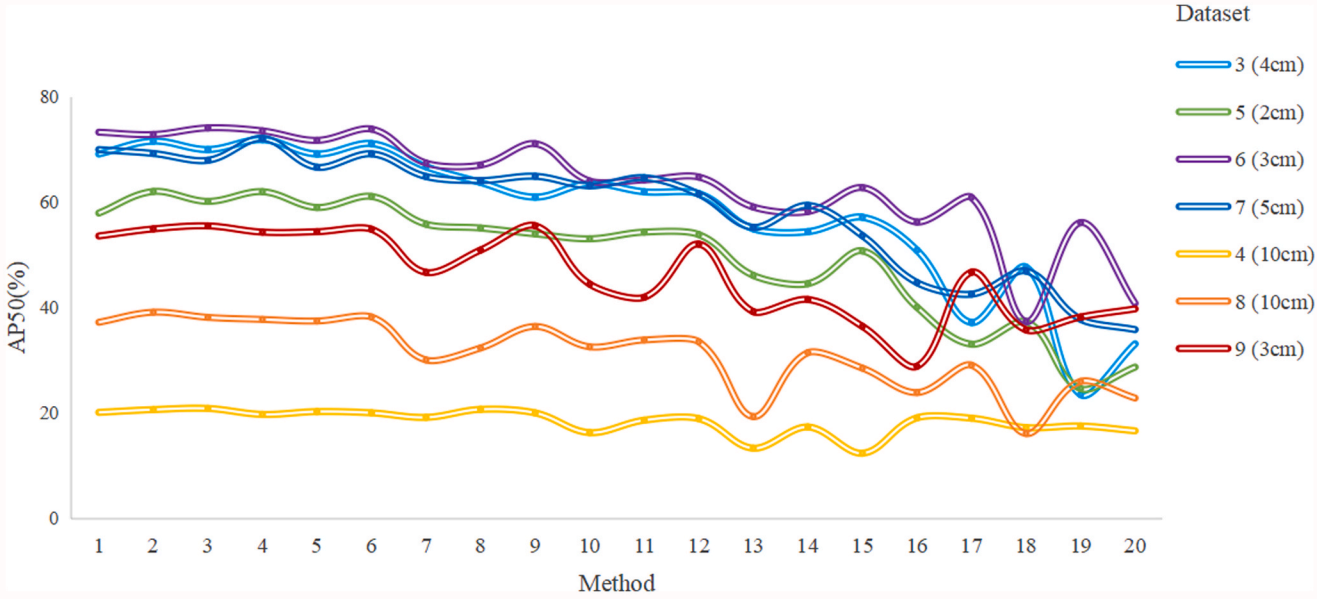

Fig. 9. The AP50 of the inferences of datasets with different resolutions.

blurring degrades the sharpness of crown textures and edges. Consequently, identifying ITC in Dataset 4 becomes challenging even with visual interpretation from human eyes, leading to a large number of mis- and under-segmentation. Meanwhile, images in Dataset 8 were affected by excessive shading caused by insufficient sunlight, which cast shadows and obscured the texture details on the crowns. Such excessive shading in images also hindered the ITD in Dataset 8, but is less significant than the blurry images in Dataset 4.

These observations emphasize the importance of collecting data under favorable weather conditions, such as avoiding conditions with insufficient or intense sunlight.

5.2. Forest conditions

The species composition and landscape characteristics vary significantly across the experimental forest sites. To assess the impact of forest conditions on ITD performance, we examined Datasets 5, 6, 10, and 11, which come from different climate zones and represent various forest types. These datasets all have good image quality with clear textures and high image resolutions that are higher than 5 cm.

The results reveal that the overall impacts of the climate zone are not as significant as expected, while site conditions, such as canopy density and species composition, play a more profound role. As shown in Fig. 13, although both Datasets 5 and 6 are from subtropical regions, their ITD performances differ considerably. Dataset 5 from a natural forest receives much lower AP50 scores in comparison with Dataset 6 from an urban forest. A closer examination of the images from these two datasets (Fig. 14) reveals that the canopy density in Dataset 5 is significantly higher in comparison with Dataset 6 and other datasets, which almost have no gaps between crowns. This dense canopy structure may explain the lower AP50 scores for Dataset 5 in comparison with other datasets with similar resolution and image quality.

Overall, the results of this study indicate that homogeneous forest stands with dense, closed canopies pose the greatest challenge for current ITD methods. In contrast, mixed forests are less challenging to image-based ITD, as the variations in crown color, texture, and size enhance the distinguishability of ITC in the images.

5.3. Method design

The performance of ITD methods is influenced by various factors in model design, including architecture choice, model training strategies, and technical enhancements.

5.3.1. Type

Both the ML- and DL-based ITD methods are investigated in this study. For the ML-based ITD from images, no commonly applied method, e.g., open code or software, is available. Among existing approaches, watershed-based segmentation has proven to be the most effective ML-based method, but it favors CHM images where height information is available. The performance of ML-based methods also depends strongly on factors including parameter settings, crown size variability, and texture heterogeneity. Significant over- and under-segmentation, as well as classification confusion, are observed in the ITD results, leading to failures in accurately identifying the number, locations, and shapes of individual tree crowns (ITCs) at most study sites. While more sophisticated model designs may improve performance, such gains often come at the expense of transferability.

The contest was open for all methodologies. Yet, all top-ranked methods are DL-based, indicating the popularity of the DL method for challenging instance segmentation tasks such as ITD. Among DL-based methods, a clear performance gap exists between the top-ranked methods in the contest and standard and large model. This gap indicates that the DL models specifically designed for the ITD task outperform large and standard instance-segmentation methods. This gap also indicates that the ITD methods in the contest represent the SOTA. Thus, the following discussions focus on the ITD methods in the contest.

The DL-based ITD methods managed to delineate quite a large proportion of dominant ITCs in the contest. Yet, the performances of the DL methods still do not meet the practical requirements of forest field investigation at this moment.

5.3.1.1. Baseline. The top six methods in the contest follow two primary architectural paradigms: query-based (Methods 1, 4, and 6) and proposal-based (Methods 2, 3, and 5) methods. When AP50 is considered, the difference between proposal- and query-based networks is insignificant, as shown in Table 6. In the case of Dataset 10 and 11 that are excluded from the training stage, the query-based Method 1 presents markedly robust performance for both unseen datasets especially Dataset 11, while the proposal-based Method 3 presents relatively robust performance only for Dataset 10. The transferability of other methods is relatively low in comparison with Method 1 and Method 3.

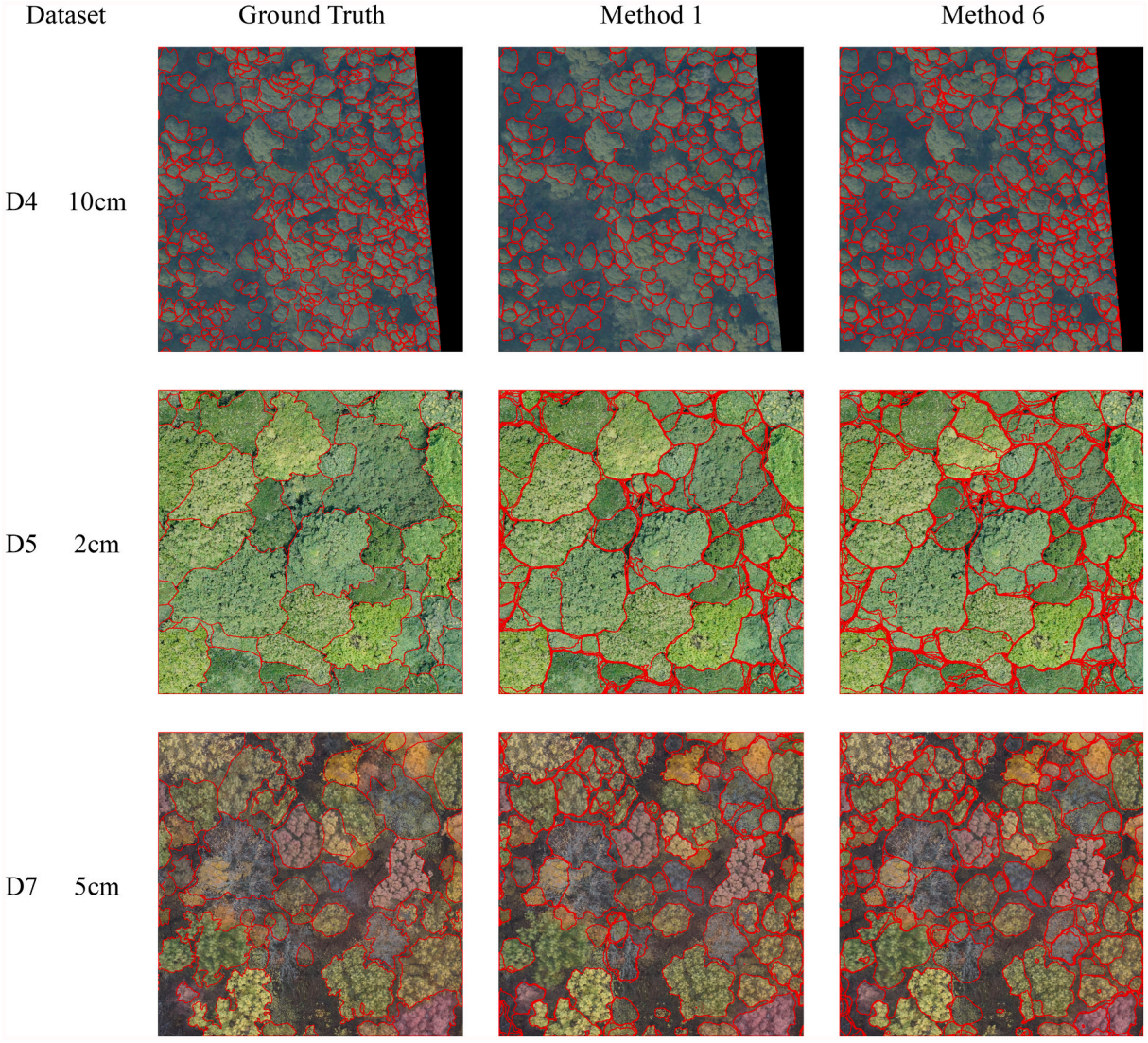

Fig. 10. The ITD results of datasets with different resolutions. The first column illustrates the ground truth. The column 2–3 represents the inference masks from Methods 1 and 6, respectively.

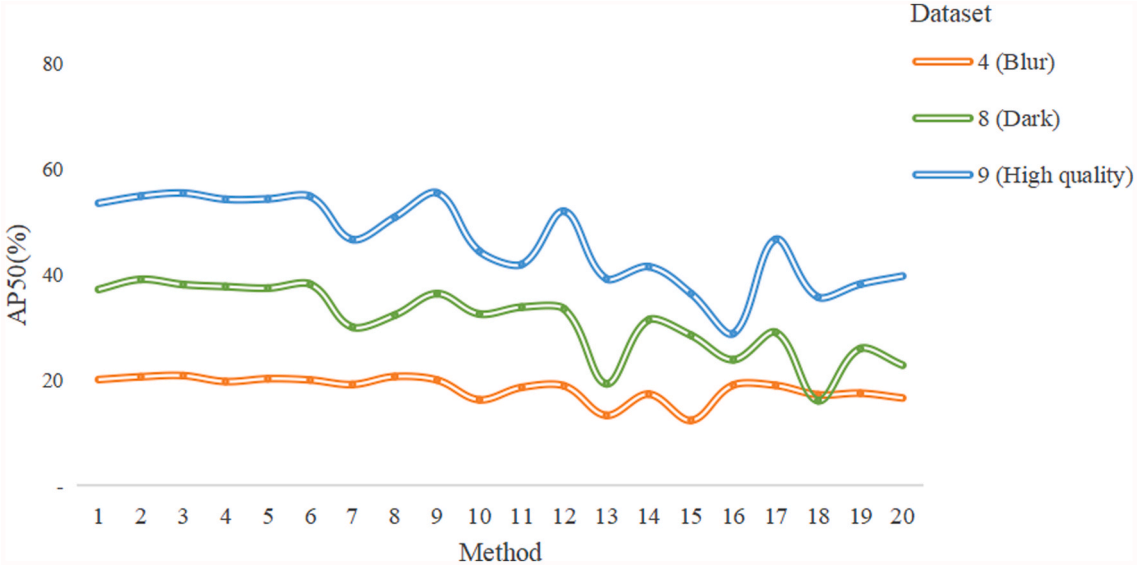

Fig. 11. The AP50 of the inferences of the datasets with different image qualities.

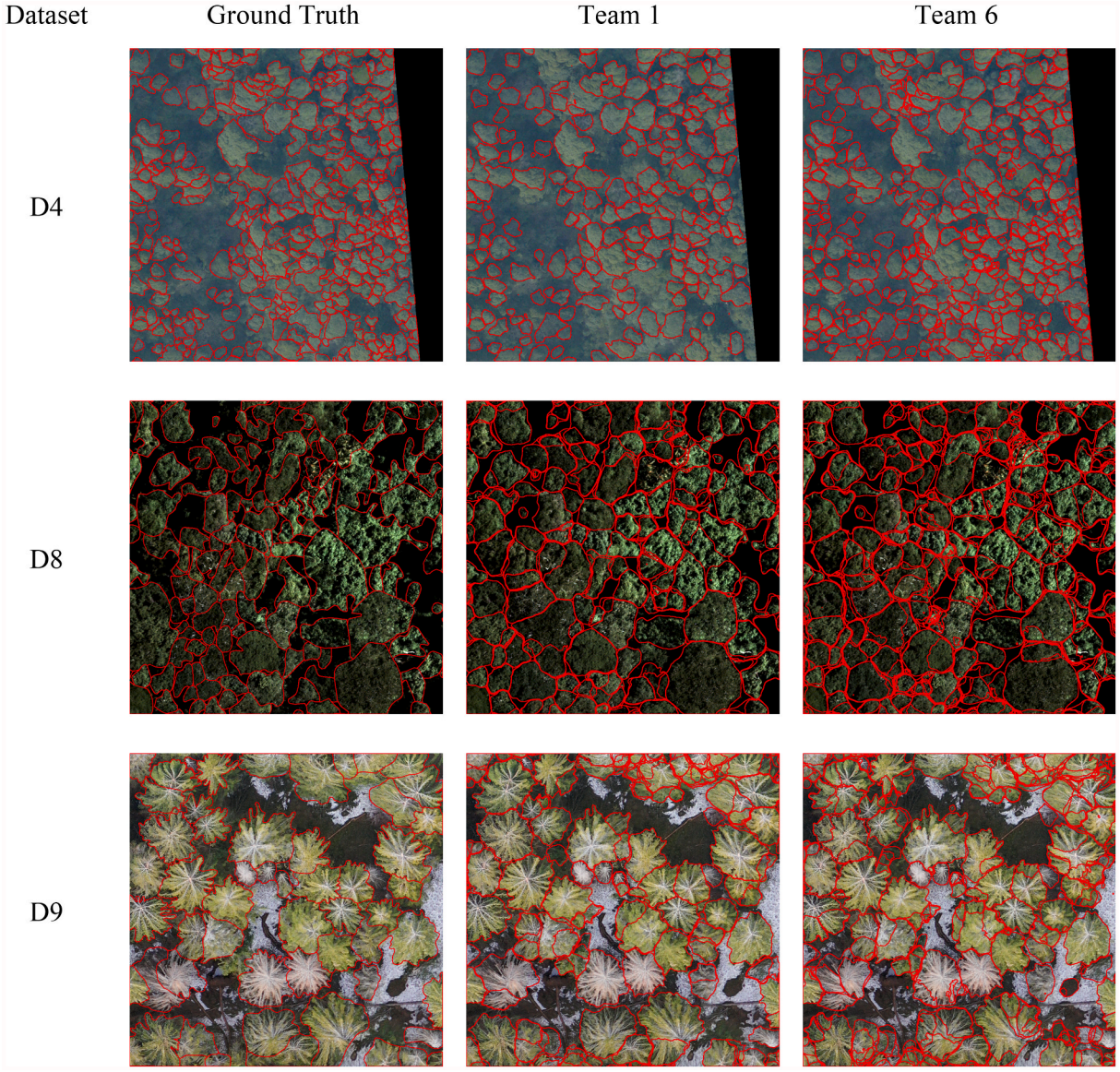

Fig. 12. The ITD results of datasets with different image qualities.

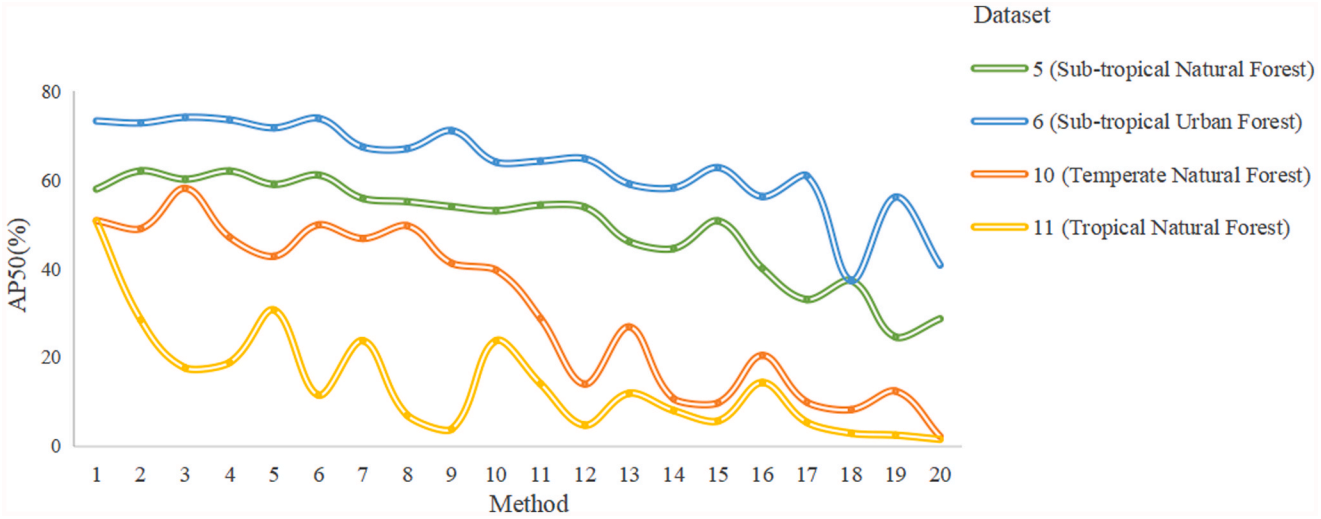

Fig. 13. The AP50 of the inferences of datasets collected from different types of forests.

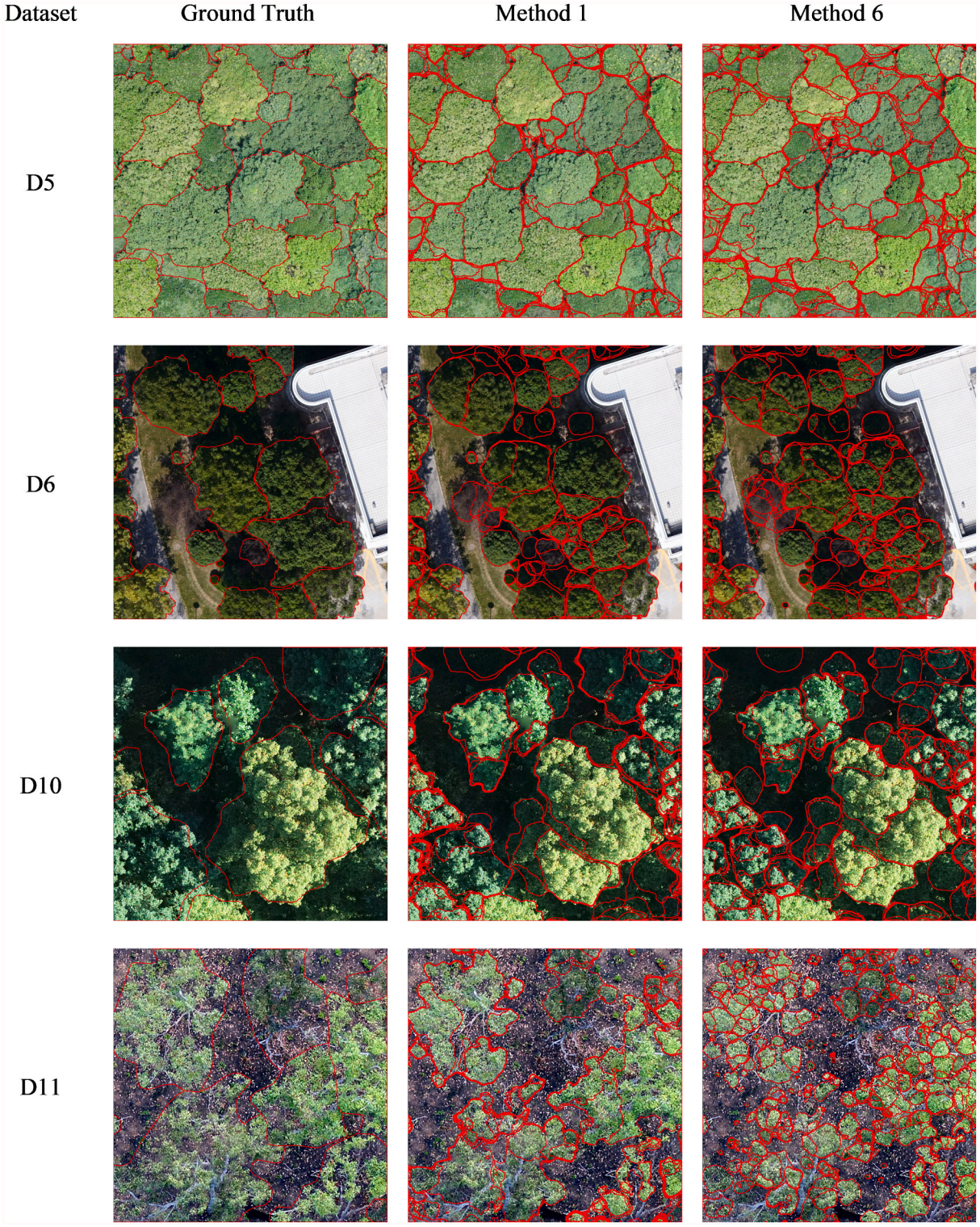

Fig. 14. The ITD results of datasets that are collected from different types of forests and with high resolutions at 2–3 cm.

However, when a stricter standard for TP is used, i.e., AP75 scores in Table 7, the proposal-based networks demonstrate better performance across most datasets, e.g., both Method 2 and Method 3 obtain the highest mean AP75 score. In contrast, the query-based methods (Method 1, 4, and 6) exhibit lower AP75 values in comparison with proposal-based Methods 2 and 3.

Regarding the generalization, both query- and proposal-based networks exhibit instability across different datasets (i.e., study sites).

5.3.1.2. Architectural modifications. In addition to the baseline architectures, specific design choices in individual components such as the backbone, initial prediction module, and head significantly influence the performance.

Among query-based methods, Methods 1, 4, and 6 are all derived from the Mask DINO baseline, with Method 6 representing the original version. The improved performance of Methods 1 and 4 over Method 6 highlights the effectiveness of their respective modifications. Method 1

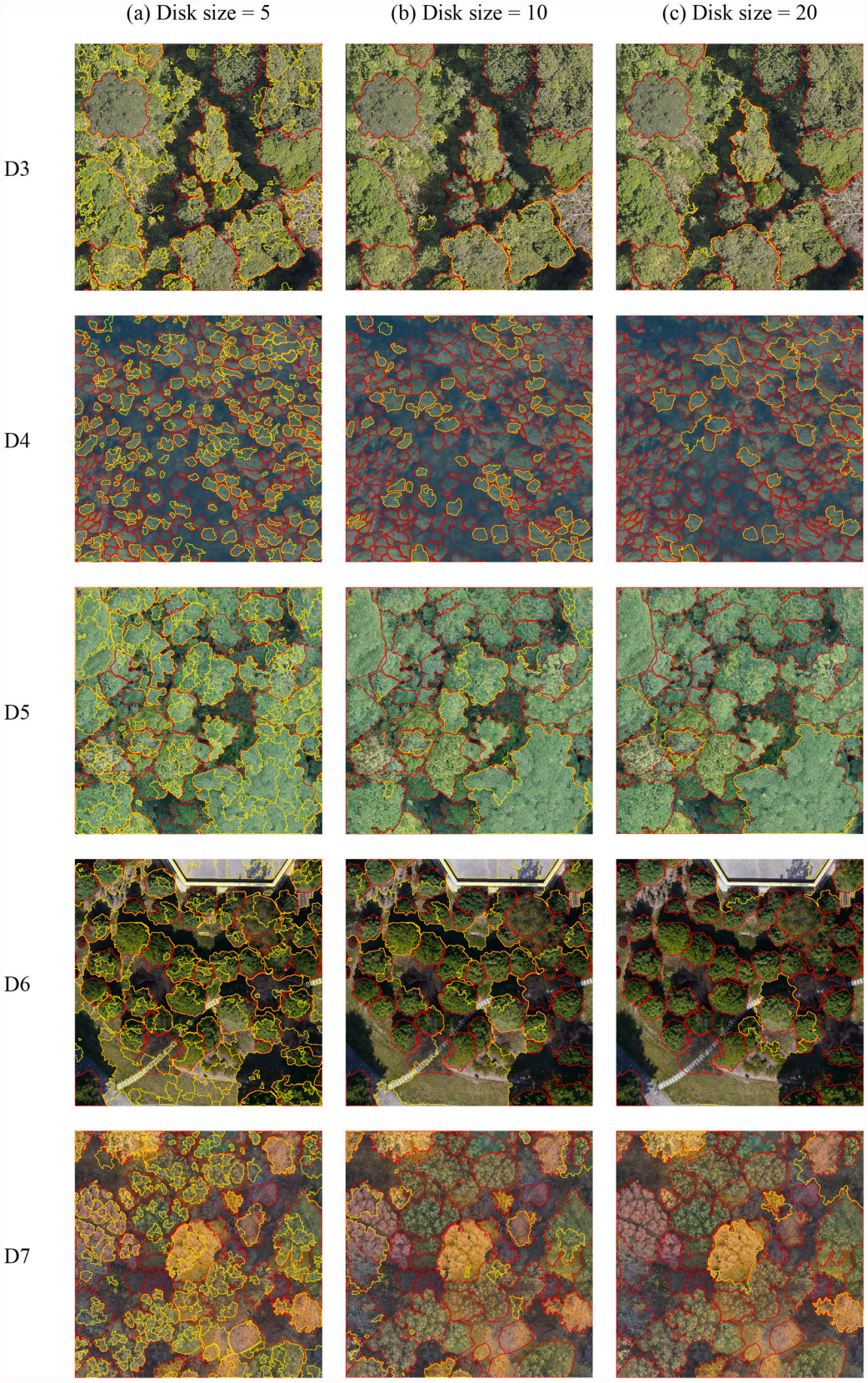

Fig. 15. The ITD results of the marker-controlled watershed segmentation with different disk sizes for the morphological pre-processing. The red and yellow polygons represent the ground truth and inference ITC masks, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

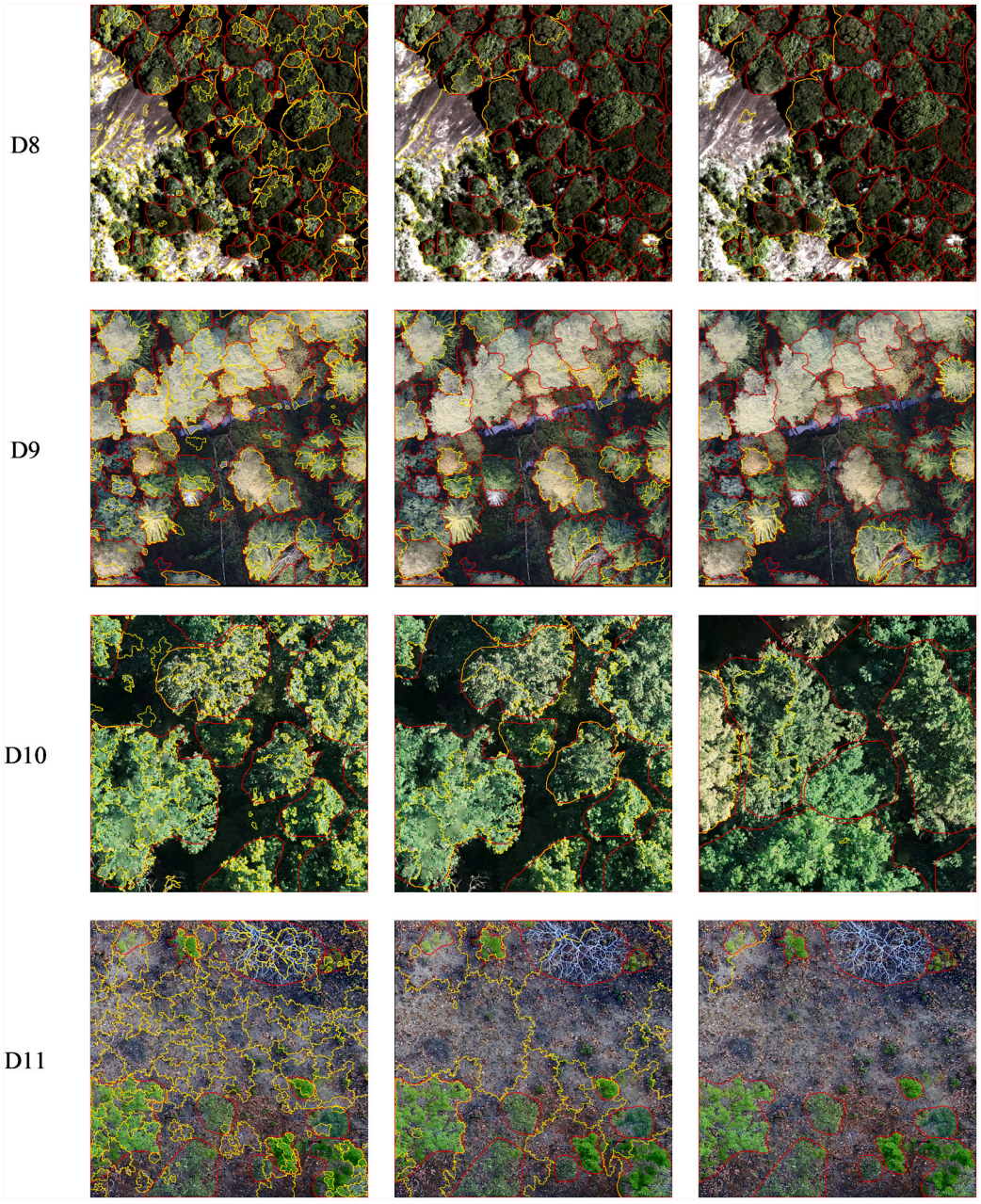

Fig. 15. (*continued*).

employs a dual Swin-L backbone with a connection mechanism, while Method 4 introduces multiple modifications across the backbone, initial prediction module, and head.

For proposal-based methods, both Methods 2 and 3 follow the HTC baseline. However, Method 2 achieved higher accuracy than Method 3 by incorporating a Mask IoU head, which enhances segmentation quality.

5.3.1.3. Technical enhancements. The technical enhancements applied in each method are summarized in Table 3. At this stage, no clear evidence supports the effectiveness of these techniques, such as test-time augmentation, in improving ITD performance. Further investigation is needed to determine their actual impacts.

5.3.1.4. Redundant prediction. Accurate ITC masks are essential for canopy parameter extraction in many practical applications. Thus, redundant predictions, where a single ITC is delineated by multiple positive masks, is popular in SOTA ITD methods. Redundant prediction improves the recall for ITD, but causes lower precision.

Redundant predictions exist in both proposal- and query-based networks. These redundant predictions stem from the excess number of queries and proposals relative to the actual number of trees. Although integrated into the networks, redundancy removal mechanisms, such as non-maximum suppression (NMS) and background classification heads, may fail to eliminate all redundant initial predictions.

5.4. Evaluation

The choice of evaluation metrics directly impacts the assessment of ITD methods. The contest adopts the AP50 as the primary evaluation metric. However, a detailed analysis reveals that, while effective for object counting, AP50 falls short in capturing finer details required for many quantitative analyses. A stricter criterion, e.g., AP75 is therefore recommended for future studies.

5.4.1. Metrics

The selection of the evaluation metric is application-dependent, namely, evaluation metrics that are suitable for one task may be inappropriate for other tasks.

The AP score is applied in this study because the conventional position-paring evaluation approach does not work for image-based ITC masks, as the treetops are not explicitly detected as with LiDAR data due to the lack of height information. AP integrates precision and recall metrics, with higher values indicating more reliable results. More specifically, the IoU-based TP identification method tanks into consideration positional, dimensional, and morphological fidelity. Thus, it evaluates the geometric agreement between the inferences and references more comprehensively, in contrast to the conventional approach that relies solely on spatial correspondence.

AP50, defined by an IoU threshold of 0.5, was used as the primary evaluation metric in the contest ranking as it remains the standard evaluation criterion in instance segmentation. This choice ensures comparability with previous studies and studies in other domains.

AP50 can be roughly equivalent to normal treetop pairing, because position paring also tolerates certain distances between the detection and reference. The typical maximum tolerable distance for position-pairing is the average crown radius in the studied area. Considering the situation of IoU = 0.5 shown in Fig. 2 (b), the distances between the gravity centers (which can be regarded as treetops) of the two crown shapes in generally do not exceed the radius of the circular crown. In general, AP50 determines the presence of objects and may be sufficient for tasks like object counting.

However, AP50 is not strict enough to evaluate the accuracy in ITD. The predicted boundary of a TP often deviates significantly from the ground truth, as visualized by the ITC masks under different IoU thresholds in Figs. 2 (b) and 4. This deviation indicates that a 0.5 IoU threshold is more appropriate for object counting rather than applications where tree-crown size and morphological fidelity are essential — such as applying segmentation results for further ecological or structural estimations (Liang et al., 2025; Shcherbacheva et al., 2024).

AP75 is a metric with sticker rule for TP identification, requiring more faithful alignment in positional, dimensional, and morphological fidelity between inferences and references in comparison with the results from AP50, as suggested in Fig. 2 (b). Applying a stricter threshold such as AP90 (IoU $\geq$ 0.9) yields more accurate TPs. However, the recall declines sharply, as illustrated in Fig. 2 (b), indicating the AP90 may be overly restrictive. IoU thresholds between 0.7 and 0.8 strike a better balance between strictness and applicability. As visualized in the ITC mask in Fig. 4, AP75 effectively balances credibility and completeness.

Given the limitations of AP50 in practical applications, AP75 or a similar metric is recommended as a suitable evaluation metric for ITD in the future, to provide a more meaningful evaluation of segmentation accuracy, where accurate crown size and morphological estimation are required. For those applications emphasizing the object count and/or location, AP50 is probably an acceptable criterion.

5.4.2. Data- vs. reality-based accuracy. The reference ITC masks are annotated from images and thus only provide information about the trees at the upper canopy layer that are visible from the images, and lack the information about the suppressed and understory trees in the secondary layer. Hence, the accuracy evaluated using the annotated reference represents the "data-based accuracy", namely, the proportion of accurately detected ITCs with respect to the total amount of ITCs captured by the images, which reveals the capability of a method in interpreting the given data.

Meanwhile, the "reality-based accuracy" is evaluated with respect to the total amount of trees standing on the ground. It indicates the performances of the method in revealing the reality of the targeted forest stand, which can be lower than the "data-based accuracy", especially the poor recall, considering the amount of suppressed small trees in a forest stand that cannot be captured by the data from viewpoints above forest canopies (Wang et al., 2019b) and the incorrect labels caused by ambiguous visual separation of ITCs with significant intersection and overlap (Allen et al., 2025).

Considering the limited accessibility of the field references due to the cost and the workload, as well as the more intuitive evaluation of method performance from the "data-based accuracy", this study focuses on evaluations with respect to the manually annotated ITC masks.

5.5. Transferability and generalization of methods

The datasets from study sites 10 and 11 were excluded from the training set for both the methods developed during the contest and the SOTA instance segmentation methods. Consequently, the segmentation results in these two study sites serve as indicators of the methods' transferability.

As shown in Fig. 8, the performances of well-designed methods are relatively similar on datasets used in the training phase. The key factor distinguishing individual methods — both those developed during the contest and the SOTA approaches — is their transferability to unseen data. Among the developed methods, Method 1 achieved better segmentation performance in Dataset 11 in comparison with other methods.

Furthermore, the zero-shot inference of SAM performs the worst in the benchmarking, demonstrating significantly weaker transferability in ITD. This finding underscores the ongoing challenges of developing universal ITD methods with high transferability across diverse forest conditions.

5.6. Outlooks

Both query- and proposal-based methods have been developed in recent years for image-based ITD applications. Despite advancements, the applicability of SOTA ITD methods in forest environments remains limited. Although some methods achieve up to 55% AP50 in certain datasets, the overall accuracy remains low, with an average of approximately 50% for AP50 and 25% for AP75.

This performance is insufficient for practical applications. When only the upper canopy trees are considered, the airborne LiDAR data provides around 80–90% detection accuracy (i.e., true positive) for upper canopy trees (i.e., dominant and codominant) depending on the complexity of forest stands, e.g., (Wang et al., 2019a). The SfM point cloud has a similar performance as the airborne LiDAR when only the upper canopy trees are considered, e.g., (Iqbal et al., 2021). It should be emphasized that, to ensure a better comparison to the outcomes of this study, the above-mentioned LiDAR and SfM performance only considers the trees on upper canopy layer. When field-measured reference is applied and all trees from upper- and suppressed-canopy layers are considered, the reality-based detection accuracy can be significantly lower for LiDAR and SfM point clouds as well, depending on the complexity of the forest stands and the proportion of the suppressed trees.

Besides the overall detection accuracy, another limitation of the pure image-based ITD approaches in comparison with the LiDAR-/SfM-based approaches is the lack of 3D tree metrics such as the crown volume, crown depth, canopy height. Nevertheless, besides the overall number of crowns in a defined area, image-based ITD is capable of providing 2D crown metrics such as the crown project area, the canopy closure, and to a certain degree, the branching architecture (i.e., texture). Moreover, an extra advantage of images is the rich spectral information. When the variation of crown spectral characteristics over time can be enhanced by multi-temporal data collection, there is a great potential for studies on tree species, phenology, and the leaf dynamics (Liang et al., 2025).

A preliminary test indicates that satellite images with 0.5 m resolution is still insufficient for ITD, especially in natural forests. This suggests that both methods and data acquisition require further enhancement to leverage the potential of Earth observation (EO) data, given the vast and

continuously growing volume of high-resolution EO optical data.

Moreover, the transferability of SOTA methods is still restricted. Experimental results suggest that transferability is the key factor differentiating the performance of individual methods. This finding indicates that the long-standing emphasis on novel and innovative aspects of model development may be overstated. Instead, greater attention should be directed towards improving adaptability that enhances the generalization across diverse forest environments and transferability to unseen forest scenes.

6. Conclusions

This study promotes research on individual tree delineation (ITD) from high-resolution non-overlapping aerial images by analyzing both newly developed, from the International ITD Contest 2024, and state-of-the-art (SOTA) instance segmentation methods, including both large and popular instance segmentation models.

The study (1) assesses the performance of ITD methods using standardized evaluation protocols and quantitative analyses, as well as the largest forest dataset available to date, (2) identifies key factors influencing algorithm performance, and (3) examines remaining challenges and outlines the possible solutions for future advancements.

Results suggest that image resolution and quality, forest conditions, and model design all directly impact the ITD performance. The deep learning (DL) methods outperform conventional machine learning (ML) methods in ITD from high-resolution imagery. With the same input imagery, the ML-based method results in more delineation mistakes, while the DL-based method generates more reliable ITD results.

While methodological innovation remains essential, greater emphasis should be placed on robustness and transferability. The preference for novel architectures should be balanced with practical applicability, ensuring that models generalize effectively across diverse environments and applications. The development of universal methods that are capable of handling multiple tasks is still a significant challenge.

From an evaluation perspective, the widely used AP50 metric is insufficient for many forest research applications, particularly those requiring precise tree crown delineation. Instead, AP75 or similar metrics should be adopted in future studies to provide a more reliable assessment of ITD accuracy. Additionally, larger and more diverse datasets are recommended to further improve ITD model development and validation.

CRediT authorship contribution statement

Xinlian Liang: Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Yinrui Wang:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation. **Jun Pan:** Writing – review & editing, Project administration, Investigation, Funding acquisition. **Janne Heiskanen:** Writing – review & editing, Data curation. **Ningning Wang:** Writing – review & editing, Data curation. **Siyu Wu:** Writing – review & editing, Data curation. **Ilja Vuorinne:** Writing – review & editing, Data curation. **Jiaojiao Tian:** Writing – review & editing, Data curation. **Jonas Troles:** Writing – review & editing, Data curation. **Myriam Cloutier:** Writing – review & editing, Data curation. **Stefano Puliti:** Writing – review & editing, Data curation. **Aishwarya Chandrasekaran:** Writing – review & editing, Data curation. **James Ball:** Writing – review & editing, Data curation. **Xiangcheng Mi:** Writing – review & editing, Data curation. **Guochun Shen:** Writing – review & editing, Data curation. **Kun Song:** Writing – review & editing, Data curation. **Guofan Shao:** Writing – review & editing, Data curation. **Rasmus Astrup:** Writing – review & editing, Data curation. **Yunsheng Wang:** Writing – review & editing, Writing – original draft, Formal analysis, Investigation. **Petri Pellikka:** Writing – review & editing, Data curation. **Mi Wang:** Writing – review & editing, Project administration, Investigation, Funding acquisition. **Jianya Gong:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that there is no known competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Science Fund program 62425102, 62442107 and 32171789.

Part of the results of this work are based on the International Individual Tree Delineation from Imagery Contest 2024. The contest would like to express gratitude to all participants, especially the top 6 prize-winning teams for sharing method descriptions, and to the data owners, both those who provided open-access datasets and those who shared unpublished data, for their contributions to the success of the contest. The list of the winners is shown on ISPRS webpage: https://www2.isprs.org/commissions/comm3/wg1/news/contest/itd-contest-2024/.

References

Abraham, N., Khan, N.M., 2019. A novel focal tversky loss function with improved attention U-net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 683–687. 10.1109/ISBI.2019.8759329.

Allen, M.J., Owen, H.J.F., Grieve, S.W.D., Lines, E.R., 2025. Manual Labelling Artificially Inflates Deep Learning-Based Segmentation Performance on RGB Images of Closed Canopy: Validation Using TLS. 10.48550/arXiv.2503.14273.

Ball, J.G.C., Hickman, S.H.M., Jackson, T.D., Koay, X.J., Hirst, J., Jay, W., Archer, M., Aubry-Kientz, M., Vincent, G., Coomes, D.A., 2023. Accurate delineation of individual tree crowns in tropical forests from aerial RGB imagery using Mask R-CNN. Remote Sens. Ecol. Conserv. 9, 641–655. https://doi.org/10.1002/rse2.332.

Cai, S., Zhang, W., Zhang, S., Yu, S., Liang, X., 2024. Branch architecture quantification of large-scale coniferous forest plots using UAV-LiDAR data. Remote Sens. Environ. 306, 114121. https://doi.org/10.1016/j.rse.2024.114121.

Cai, Z., Vasconcelos, N., 2021. Cascade R-CNN: high quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1483–1498. https://doi.org/10.1109/TPAMI.2019.2956516.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 213–229.

Chandrasekaran, A., Hupy, J.P., Shao, G., 2024. Multi-scale mapping and analysis of broadleaf species distribution using remotely piloted aircraft and satellite imagery. Remote Sens. (Basel) 16, 4809. https://doi.org/10.3390/rs16244809.

Chemura, A., Van Duren, I., Van Leeuwen, L.M., 2015. Determination of the age of oil palm from crown projection area detected from WorldView-2 multispectral remote sensing data: the case of Ejisu-Juaben district, Ghana. ISPRS J. Photogramm. Remote Sens. 100, 118–127. https://doi.org/10.1016/j.isprsjprs.2014.07.013.

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z., 2020. Dynamic convolution: attention over convolution kernels. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Chen, J., Liang, X., Liu, Z., Gong, W., Chen, Y., Hyyppä, J., Kukko, A., Wang, Y., 2024. Tree species recognition from close-range sensing: A review. Remote Sens. of Environ. 313, 114337. https://doi.org/10.1016/j.rse.2024.114337.

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1299.

Cloutier, M., Germain, M., Laliberté, E., 2024. Influence of temperate forest autumn leaf phenology on segmentation of tree species from UAV imagery using deep learning. Remote Sens. Environ. 311, 114283. https://doi.org/10.1016/j.rse.2024.114283.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). Presented at the 2017 IEEE International Conference on Computer Vision (ICCV), pp. 764–773. https://doi.org/10.1109/ICCV.2017.89.

Dersch, S., Schöttl, A., Krzystek, P., Heurich, M., 2024. Semi-supervised multi-class tree crown delineation using aerial multispectral imagery and lidar data. ISPRS J. Photogramm. Remote Sens. 216, 154–167. https://doi.org/10.1016/j.isprsjprs.2024.07.032.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Gan, Y., Wang, Q., Iio, A., 2023. Tree crown detection and delineation in a temperate deciduous forest from UAV RGB imagery using deep learning approaches: effects of spatial resolution and species characteristics. Remote Sens. (Basel) 15. https://doi.org/10.3390/rs15030778.

Goodman, R.C., Phillips, O.L., Baker, T.R., 2014. The importance of crown dimensions to improve tropical tree biomass estimates. Ecol. Appl. 24, 680–698. https://doi.org/10.1890/13-0070.1.

Gougeon, F.A., 1995. A crown-following approach to the automatic delineation of individual tree crowns in high spatial resolution aerial images. Can. J. Remote. Sens. 21, 274–284. https://doi.org/10.1080/07038992.1995.10874622.

Harikumar, A., Liang, X., Bovolo, F., Bruzzone, L., 2022. Void-volume-based stem geometric modeling and branch-knot localization in terrestrial laser scanning data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 3024–3040.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. 10.1109/ICCV.2017.322.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X., 2019. Mask scoring R-CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Hyyppä, E., Hyyppä, J., Hakala, T., Kukko, A., Wulder, M.A., White, J.C., Pyörälä, J., Yu, X., Wang, Y., Virtanen, J.-P., Pohjavirta, O., Liang, X., Holopainen, M., Kaartinen, H., 2020a. Under-canopy UAV laser scanning for accurate forest field measurements. ISPRS J. Photogramm. Remote Sens. 164, 41–60. https://doi.org/10.1016/j.isprsjprs.2020.03.021.

Hyyppä, E., Kukko, A., Kaijaluoto, R., White, J.C., Wulder, M.A., Pyörälä, J., Liang, X., Yu, X., Wang, Y., Kaartinen, H., Virtanen, J.-P., Hyyppä, J., 2020b. Accurate derivation of stem curve and volume using backpack mobile laser scanning. ISPRS J. Photogramm. Remote Sens. 161, 246–262. https://doi.org/10.1016/j.isprsjprs.2020.01.018.

Hyyppa, J., Kelle, O., Lehikoinen, M., Inkinen, M., 2001. A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. IEEE Trans. Geosci. Remote Sens. 39, 969–975. https://doi.org/10.1109/36.921414.

Iqbal, I.A., Osborn, J., Stone, C., Lucieer, A., 2021. A Comparison of ALS and dense photogrammetric point clouds for individual tree detection in radiata pine plantations. Remote Sens. (Basel) 13, 3536. https://doi.org/10.3390/rs13173536.

Jansen, A.J., Nicholson, J.D., Esparon, A., Whiteside, T., Welch, M., Tunstill, M., Paramjyothi, H., Gadhiraju, V., van Bodegraven, S., Bartolo, R.E., 2023. Deep learning with northern australian savanna tree species: a novel dataset. Data 8. https://doi.org/10.3390/data8020044.

Jing, L., Hu, B., Noland, T., Li, J., 2012. An individual tree crown delineation method based on multi-scale segmentation of imagery. ISPRS J. Photogramm. Remote Sens. 70, 88–98. https://doi.org/10.1016/j.isprsjprs.2012.04.003.

Kaartinen, H., Hyyppä, J., Yu, X., Vastaranta, M., Hyyppä, H., Kukko, A., Holopainen, M., Heipke, C., Hirschmugl, M., Morsdorf, F., Næsset, E., Pitkänen, J., Popescu, S., Solberg, S., Wolf, B.M., Wu, J.-C., 2012. An international comparison of individual tree detection and extraction using airborne laser scanning. Remote Sens. (Basel) 4, 950–974. https://doi.org/10.3390/rs4040950.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P., Girshick, R., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026.

Kotaridis, I., Lazaridou, M., 2021. Remote sensing image segmentation advances: a meta-analysis. ISPRS J. Photogramm. Remote Sens. 173, 309–322. https://doi.org/10.1016/j.isprsjprs.2021.01.020.

Lee, C.K.F., Song, G., Muller-Landau, H.C., Wu, S., Wright, S.J., Cushman, K.C., Araujo, R.F., Bohlman, S., Zhao, Y., Lin, Z., Sun, Z., Cheng, P.C.Y., Ng, M.-K.-P., Wu, J., 2023. Cost-effective and accurate monitoring of flowering across multiple tropical tree species over two years with a time series of high-resolution drone imagery and deep learning. ISPRS J. Photogramm. Remote Sens. 201, 92–103. https://doi.org/10.1016/j.isprsjprs.2023.05.022.

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L., 2022. DN-DETR: accelerate DETR training by introducing query DeNoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13619–13627.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.-Y., 2023. Mask DINO: towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3041–3050.

Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H., 2022a. CBNet: a composite backbone network architecture for object detection. IEEE Trans. Image Process. 31, 6893–6906. https://doi.org/10.1109/TIP.2022.3216771.

Liang, X., Chen, J., Gong, W., Puttonen, E., Wang, Y., 2025. Influence of data and methods on high-resolution imagery-based tree species recognition considering phenology: the case of temperate forests. Remote Sens. Environ. 323, 114654. https://doi.org/10.1016/j.rse.2025.114654.

Liang, X., Kukko, A., Balenović, I., Saarinen, N., Junttila, S., Kankare, V., Holopainen, M., Mokroš, M., Surový, P., Kaartinen, H., Jurjević, L., Honkavaara, E., Näsi, R., Liu, J., Hollaus, M., Tian, J., Yu, X., Pan, J., Cai, S., Virtanen, J.-P., Wang, Y., Hyyppä, J., 2022b. Close-range remote sensing of forests: the state of the art, challenges, and opportunities for systems and data acquisitions. IEEE Geosci. Remote Sens. Mag. 10, 32–71. https://doi.org/10.1109/MGRS.2022.3168135.

Liang, X., Qi, H., Deng, X., Chen, J., Cai, S., Zhang, Q., Wang, Y., Kukko, A., Hyyppä, J., 2024a. ForestSemantic: a dataset for semantic learning of forest from close-range sensing. Geo-spat. Inf. Sci. https://doi.org/10.1080/10095020.2024.2313325.

Liang, X., Wang, Y., Pan, J., Wang, M., Yang, J., Gong, J., 2024b. The ISPRS international contest on individual tree crown segmentation using high-resolution images and the initial findings. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. https://doi.org/10.5194/isprs-archives-XLVIII-3-2024-637-2024.

Liang, X., Yao, H., Qi, H., Wang, X., 2024c. Forest in situ observations through a fully automated under-canopy unmanned aerial vehicle. Geo-spatial Inf. Sci. https://doi.org/10.1080/10095020.2024.2322765.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 740–755.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986.

Lowman, M.D., Rinker, H.B., 2004. Introduction. pp. xxi–xxiii. In: Lowman, M.D., Rinker, H.B. (Eds.), Forest Canopies (second Edition), Physiological Ecology. Academic Press, San Diego. https://doi.org/10.1016/B978-012457553-0/50003-4.

Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., Horvath, P., 2020. Test-time augmentation for deep learning-based cell segmentation on microscopy images. Sci. Rep. 10, 5068. https://doi.org/10.1038/s41598-020-61808-3.

Puliti, S., Astrup, R., 2022. Automatic detection of snow breakage at single tree level using YOLOv5 applied to UAV imagery. Int. J. Appl. Earth Obs. Geoinf. 112, 102946. https://doi.org/10.1016/j.jag.2022.102946.

Pyörälä, J., Liang, X., Saarinen, N., Kankare, V., Wang, Y., Holopainen, M., Hyyppä, J., Vastaranta, M., 2018. Assessing branching structure for biomass and wood quality estimation using terrestrial laser scanning point clouds. Can. J. Remote. Sens. 44, 462–475. https://doi.org/10.1080/07038992.2018.1557040.

Qi, H., Liang, X., 2024. Automated first-order tree branch modeling at plot- and individual-tree-levels from close-range sensing for silviculture and forest ecology. IEEE Trans. Geosci. Remote Sens. 62, 1–21. https://doi.org/10.1109/TGRS.2024.3443259.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: a metric and a loss for bounding box regression. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Ross, T.-Y., Dollár, G., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2980–2988.

Safonova, A., Hamad, Y., Dmitriev, E., Georgiev, G., Trenkin, V., Georgieva, M., Dimitrov, S., Iliev, M., 2021. Individual tree crown delineation for the species classification and assessment of vital status of forest stands from UAV images. Drones 5. https://doi.org/10.3390/drones5030077.

Sani-Mohammed, A., Yao, W., Heurich, M., 2022. Instance segmentation of standing dead trees in dense forest from aerial imagery using deep learning. ISPRS Open J. Photogram. Rem. Sens. 6, 100024. https://doi.org/10.1016/j.ophoto.2022.100024.

Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J., 2020. When and why test-time augmentation works. arXiv preprint arXiv:2011.11156 1, 4.

Shcherbacheva, A., Campos, M.B., Wang, Y., Liang, X., Kukko, A., Hyyppä, J., Junttila, S., Lintunen, A., Korpela, I., Puttonen, E., 2024. A study of annual tree-wise LiDAR intensity patterns of boreal species observed using a hyper-temporal laser scanning time series. Remote Sens. Environ. 305, 114083. https://doi.org/10.1016/j.rse.2024.114083.

Sun, Y., Li, Z., He, H., Guo, L., Zhang, X., Xin, Q., 2022. Counting trees in a subtropical mega city using the instance segmentation method. Int. J. Appl. Earth Obs. Geoinf. 106, 102662. https://doi.org/10.1016/j.jag.2021.102662.

Spies, Thomas A, 1998. Forest structure: a key to the ecosystem. In: S. Proceedings of a Workshop on Structure, Process, and Diversit. pp. 34–39.

Troles, J., Schmid, U., Fan, W., Tian, J., 2024. BAMFORESTS: bamberg benchmark forest dataset of individual tree crowns in very-high-resolution UAV images. Remote Sens. (Basel) 16. https://doi.org/10.3390/rs16111935.

Vauhkonen, J., Ene, L., Gupta, S., Heinzel, J., Holmgren, J., Pitkänen, J., Solberg, S., Wang, Y., Weinacker, H., Hauglin, K.M., Lien, V., Packalén, P., Gobakken, T., Koch, B., Næsset, E., Tokola, T., Maltamo, M., 2011. Comparative testing of single-tree detection algorithms under different types of forest. Forest.: Int. J. Forest Res. 85, 27–40. https://doi.org/10.1093/forestry/cpr051.

Wang, L., 2010. A multi-scale approach for delineating individual tree crowns with very high resolution imagery. Photogram. Eng. Remote Sens. 76, 371–378.
Wang, L., 2003. Object-based methods for individual tree identification and tree species classification from high-spatial resolution imagery (Ph.D.). ProQuest Dissertations and Theses. University of California, Berkeley, United States – California.
Wang, L., Gong, P., Biging, G.S., 2004. Individual tree-crown delineation and treetop detection in high-spatial-resolution aerial imagery. Photogram. Eng. Rem. Sens. 70, 351–357. https://doi.org/10.14358/PERS.70.3.351.
Wang, Y., Hyyppä, J., Liang, X., Kaartinen, H., Yu, X., Lindberg, E., Holmgren, J., Qin, Y., Mallet, C., Ferraz, A., Torabzadeh, H., Morsdorf, F., Zhu, L., Liu, J., Alho, P., 2016. International benchmarking of the individual tree detection methods for modeling 3-D canopy structure for silviculture and forest ecology using airborne laser scanning. IEEE Trans. Geosci. Remote Sens. 54, 5011–5027. https://doi.org/10.1109/TGRS.2016.2543225.
Wang, Y., Kukko, A., Hyyppä, E., Hakala, T., Pyörälä, J., Lehtomäki, M., El Issaoui, A., Yu, X., Kaartinen, H., Liang, X., Hyyppä, J., 2021. Seamless integration of above- and under-canopy unmanned aerial vehicle laser scanning for forest investigation. For. Ecosyst. 8, 10. https://doi.org/10.1186/s40663-021-00290-3.
Wang, Y., Lehtomäki, M., Liang, X., Pyörälä, J., Kukko, A., Jaakkola, A., Liu, J., Feng, Z., Chen, R., Hyyppä, J., 2019a. Is field-measured tree height as reliable as believed – a comparison study of tree height estimates from field measurement, airborne laser scanning and terrestrial laser scanning in a boreal forest. ISPRS J. Photogramm. Remote Sens. 147, 132–145. https://doi.org/10.1016/j.isprsjprs.2018.11.008.
Wang, Y., Pyörälä, J., Liang, X., Lehtomäki, M., Kukko, A., Yu, X., Kaartinen, H., Hyyppä, J., 2019b. In situ biomass estimation at tree and plot levels: what did data record and what did algorithms derive from terrestrial and aerial point clouds in boreal forest. Remote Sens. Environ. 232, 111309. https://doi.org/10.1016/j.rse.2019.111309.
Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16133–16142.
Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2.
Xie, Y., Wang, Y., Sun, Z., Liang, R., Ding, Z., Wang, B., Huang, S., Sun, Y., 2024. Instance segmentation and stand-scale forest mapping based on UAV images derived RGB and CHM. Comput. Electron. Agric. 220, 108878. https://doi.org/10.1016/j.compag.2024.108878.
Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y., 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection.
Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N., 2021. VarifocalNet: an IoU-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8514–8523.
Zhang, Q., Cai, S., Liang, X., 2024. Individual tree segmentation in occluded complex forest stands through ellipsoid directional searching and point compensation. For. Ecosyst. https://doi.org/10.1016/j.fecs.2024.100238.
Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J., 2024. DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965–16974.
Zhu, F., Chen, Z., Li, H., Shi, Q., Liu, X., 2024. CEDAnet: individual tree segmentation in dense orchard via context enhancement and density prior. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 7040–7051. https://doi.org/10.1109/JSTARS.2024.3378167.
Zhu, X., Hu, H., Lin, S., Dai, J., 2018. Deformable ConvNets v2: more deformable. Bet. Results, 10.48550/ARXIV.1811.11168.
Zong, Z., Song, G., Liu, Y., 2023. DETRs with collaborative hybrid assignments training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6748–6758.