ISPRS Open Journal of Photogrammetry and Remote Sensing 8 (2023) 100037

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing

Towards complete tree crown delineation by instance segmentation with Mask R–CNN and DETR using UAV-based multispectral imagery and lidar data

S. Dersch [a,b,f,*], A. Schöttl [e,f], P. Krzystek [a,f], M. Heurich [b,c,d]

[a] Dept. of Geoinformatics, Munich University of Applied Sciences, 80333, Munich, Germany
[b] Faculty of Environment and Natural Resources, University of Freiburg, Germany
[c] Bavarian Forest National Park, Dept. of Visitor Management and National Park Monitoring, 94481, Grafenau, Germany
[d] Institute of Forestry and Wildlife Management, Inland Norway University of Applied Science, NO-2480, Koppang, Norway
[e] Dept. of Electrical Engineering and Information Technology, Munich University of Applied Sciences, 80335, Munich, Germany
[f] Institute for Applications of Machine Learning and Intelligent Systems, Munich University of Applied Sciences, 80335, Munich, Germany

ARTICLE INFO

Keywords:
Single tree delineation
Instance segmentation
Multispectral imagery
Lidar

ABSTRACT

Precise single tree delineation allows for a more reliable determination of essential parameters such as tree species, height and vitality. Methods of instance segmentation are powerful neural networks for detecting and segmenting single objects and have the potential to push the accuracy of tree segmentation methods to a new level. In this study, two instance segmentation methods, Mask R–CNN and DETR, were applied to precisely delineate single tree crowns using multispectral images and images generated from UAV lidar data. The study area was in Bavaria, 35 km north of Munich (Germany), comprising a mixed forest stand of around 7 ha characterised mainly by Norway spruce (*Picea abies*) and large groups of European beeches (*Fagus sylvatica*) with 181–236 trees per ha. The data set, consisting of multispectral images and lidar data, was acquired using a Micasense RedEdge-MX dual camera system and a Riegl miniVUX-1UAV lidar scanner, both mounted on a hexacopter (DJI Matrice 600 Pro). At an altitude of approximately 85 m, two flight missions were conducted at an airspeed of 5 m/s, leading to a ground resolution of 5 cm and a lidar point density of 560 points/m^2. In total, 1408 trees were marked by visual interpretation of the remote sensing data for training and validating the classifiers. Additionally, 125 trees were surveyed by tacheometric means used to test the optimized neural networks. The evaluations showed that segmentation using only multispectral imagery performed slightly better than with images generated from lidar data. In terms of F1 score, Mask R–CNN with color infrared (CIR) images achieved 92% in coniferous, 85% in deciduous and 83% in mixed stands. Compared to the images generated by lidar data, these scores are the same for coniferous and slightly worse for deciduous and mixed plots, by 4% and 2%, respectively. DETR with CIR images achieved 90% in coniferous, 81% in deciduous and 84% in mixed stands. These scores were 2%, 1%, and 2% worse, respectively, compared to the lidar data images in the same test areas. Interestingly, four conventional segmentation methods performed significantly worse than CIR-based and lidar-based instance segmentations. Additionally, the results revealed that tree crowns were more accurately segmented by instance segmentation. All in all, the results highlight the practical potential of the two deep learning-based tree segmentation methods, especially in comparison to baseline methods.

1. Introduction

Forests are important for our ecosystems, providing and regulating services for human livelihoods. This includes providing raw wood and fresh water, as well as regulating the climate and air quality (Reid et al., 2005). Furthermore, forests play an important role in the global carbon cycle and thus are crucial for the mitigation of global change (Seidl et al., 2014). In recent years they have faced more frequent and severe disturbances that might exaggerate their resilience (Lindner et al., 2010; Thom and Seidl, 2016). Due to the crucial services forest ecosystems

* Corresponding author. Dept. of Geoinformatics, Munich University of Applied Sciences, 80333, Munich, Germany.
E-mail addresses: sebastian.dersch@hm.edu (S. Dersch), alfred.schoettl@hm.edu (A. Schöttl), peter.krzystek@hm.edu (P. Krzystek), marco.heurich@npv-bw.bayern.de (M. Heurich).

https://doi.org/10.1016/j.ophoto.2023.100037
Received 6 December 2022; Received in revised form 19 April 2023; Accepted 19 April 2023
Available online 5 May 2023
2667-3932/© 2023 The Authors. Published by Elsevier B.V. on behalf of International Society of Photogrammetry and Remote Sensing (isprs). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

provide, an accurate monitoring system is of the utmost importance. So far, traditional forest inventories have been based on so-called sample plots (less than 1% of the total forest area). On these small forest plots, individual tree attributes are surveyed by field measurements and used for the calculation of statistical indicators for larger management units (e.g. forest districts) using statistical models (Heurich, 2006). These techniques are very time-consuming because of the enormous amount of human resources involved. The accuracy of inventory parameters for large spatial units is high (i.e. biomass, growing stock); however, the predictions for forest stands show large variability due to the low number of sampling units. Instead, remote sensing techniques can provide accurate solutions for seamless calculation of relevant forest inventory parameters.

Remote sensing instruments such as lidar and multispectral cameras have been widely used for data acquisition in forest areas (White et al., 2016). Recently, sensors have been miniaturised for unmanned aerial vehicle (UAV) applications. Based on the highly precise sensor data, forest structure parameters can be determined with either area-based or tree-level approaches. Recently, Latifi et al. (2015) demonstrated that single tree-based algorithms can reliably estimate forest structure variables that are useable for forest inventories, studies on biodiversity and growth models. Here, techniques for segmentation and detection of single trees come into play that provide tree attributes such as position, height, crown volume and biomass (Yao et al., 2012). However, an imprecise crown delineation with under- and over-segmentation reduces the quality of the obtained forest structure parameters (Yu et al., 2010).

A plethora of single tree approaches can be found in the literature (Vauhkonen et al., 2011). In addition to conventional methods like watershed segmentation of a digital surface model (DSM) or truly three-dimensional (3D) approaches using lidar point clouds, new deep learning based methods have been proposed that take advantage of the potential of neural networks and, thereby, outperform conventional machine learning methods (Ren et al., 2017; Ronneberger et al., 2015; He et al., 2017). These new methods are advantageous because they extract features fully automatically during the training process (LeCun et al., 2015). Furthermore, approaches using the mechanism of self-attention (Vaswani et al., 2017) have gained momentum. The enlarged receptive field offers an advantage in detecting and segmenting large objects, and it has found broad interest in the computer vision community (Carion et al., 2020).

Aerial forest surveys are usually conducted with airplanes or helicopters. The use of UAVs has thus far been limited to smaller areas, because without getting special permission, drones can only be used in visual flight. However, due to their low flying altitude, drone flights can achieve extremely high resolutions in recorded data, and thus they have the capability recording forest structures in great detail. Copter drones equipped with powerful batteries or vertical take-off and landing-UAVs are particularly suitable for data acquisition over small forest areas or control plots, which should be recorded with a high resolution. Because of the new potential applications, multispectral sensors and lidar instruments have been specialized for UAV applications. Today, a plethora of sensors are available. Among them are the well-known multispectral cameras, such as RedEdge-MX (2020) and Sentera6X (2019), and lidar sensors such as YellowScanVoyager (2022) and RIEGL (2020). If mounted to a drone, these sensors enable a highly detailed recording of the forest canopy, which is key for a detailed reconstruction of single trees.

The contributions of this study are as follows. First, we introduce a novel transformer-based network for individual tree delineation from multispectral imagery and high-resolution lidar data that are simultaneously recorded by a drone. Furthermore, we show how this type of network is adapted for this task and discuss its challenges and benefits. The study area was in Bavaria, 35 km north of Munich, Germany. It comprises deciduous, coniferous and mixed forest areas with a stem density of around 230 stems/ha. Second, we optimize the enclosing polygons of delineated single trees using instance segmentation; thus, we reduce the proportion of neighboring trees in the delineated tree segment. We proved that instance segmentation using optical imagery slightly outperforms lidar-based segmentation. Moreover, the performance of tree crown segmentation by instance segmentation is clearly superior to conventional baseline methods. Regarding detection rate and quality of tree crown delineation, all four baseline methods showed worse performances.

1.1. Related work

In the following, we subdivide methods for single tree detection into three categories. The first category comprises raster-based methods using lidar-based canopy height models (CHMs). These methods were developed from the 1990s to the 2010s and are known for working well on dominant trees and failing in detecting understory trees (Heurich, 2008). For instance, the approach of Silva et al. (2016) generates tree segments using local maxima from CHM and centroidal Voronoi tessellation. The research site comprised an open canopy longleaf pine (*Pinus palustris*) forest area located in southwestern Georgia (USA). The authors reported an overall F1 score of 83% on 15 test plots. However, tree detection accuracy is greatly affected by the parameters treetop window size that defines the size of the sliding window for searching treetops in the CHM and the smoothing window size that defines the window size for smoothing the CHM. Dalponte and Coomes (2016) detected single trees using a region-growing algorithm in a mixed forest located in the Italian Alps. The research site, at an altitude of 900–2200 m above sea level, is dominated by smaller Norway spruce trees. The authors reported a mean detection rate of 30.6% for 47 circular validation plots 15 m in diameter. The method clearly performed better on larger trees than on smaller ones.

The methods in the second category work on the entire lidar point cloud. Therefore, in principle, they are not limited to a surface model and can detect trees in the understory. Note that the lidar point cloud requires sufficient point density to reflect structures below the tree canopy. Reitberger et al. (2009) segmented individual trees using full-waveform lidar data with a graph-based method called normalized cut located in the Bavarian Forest National Park (Germany). The study area is characterised by Norway spruce and European beech (*Fagus sylvatica*). For 12 test plots, they reported F1 scores of up to 88%, with significant improvements in the intermediate and upper forest layers. However, there were only minor improvements in the lower layer due to low point density below the canopy.

In a study from Krzystek et al. (2020), this method was successfully applied in a large area in the Bavarian Forest (Germany) and Šumava National Park (Czech Republic). The research sites have a stem density of around 540 stems/ha and are characterised by Norway spruce and European beech. The researchers reported F1 scores up to 87% for coniferous and 72% for deciduous forest stands. The results are slightly worse than those from (Reitberger et al., 2009), which is due to the different forest structures in the test data. The most sensitive threshold parameter called $NCut_{max}$ was observed to depend on the tree species (deciduous and coniferous). Therefore, it must be calibrated and applied separately. In another study, Dersch et al. (2021) demonstrated an adaptive stopping criterion for the normalized cut method that decouples the most sensitive parameter $NCut_{max}$ using tree positions calculated by an integrated tree stem detection technique. Tree stems are automatically located by vertical lines based on a three-stage hierarchical classification process. The study area consisting of mixed forest stands (e.g. Norway spruce and European beech) was located in Styria, Austria, and had a very high stem density of around 1000 stems/ha. A remarkable improvement in comparison to the vanilla normalized cut of up to 15% in terms of F1 score was achieved in one mixed and two deciduous forest plots. However, a number of tree stems could not be detected due to insufficient point density in the lower parts of the forest.

The approach suggested by Li et al. (2012) segmented individual trees in a mixed coniferous forest in the Sierra Nevada Mountains of

California (USA), which is dominated by white fir (*Abies concolor*), ponderosa pine (*Pinus ponderosa*) and black oak (*Quercus kelloggii*), among others. The researchers used a top-down growing approach, which selects starting points as the highest points within a predefined search radius, and the resulting F1 score was 90%. The researchers reported issues with segmenting large deciduous trees with complex crowns and elongated branches. The bottom-up segmentation method used by Strîmbu and Strîmbu (2015) was based on a weighted graph that quantified topological relationships of tree crown components built from hierarchical structures of lidar data. The research site was located in Louisiana (USA) and comprised mainly coniferous stands (e.g. *Pinus taeda*, *Pinus echinata*), and the resulting tree detection rate in the upper forest layers was 97% for regular tree structures and 89% for complex tree structures. However, due to the abstract parameters of the algorithm, it was difficult to find the correct parameterization, and they needed to be determined for new forest areas.

In recent years, deep learning-based methods have gained momentum and are represented in many research realms with the prospect of better accuracies. In remote sensing, multispectral image data offers a suitable data basis for the successful application of convolutional neural networks (CNNs). They can be applied to a variety of problems, including classifying entire images or detecting and segmenting objects inside images (Kattenborn et al., 2021). The latter method is referred to as instance segmentation. There are mainly two types of established object detectors. First, there are single-stage detectors, such as single-shot multibox detector (SSD) (Liu et al., 2016), you only look once (YOLO) (Redmon and Farhadi, 2016) and RetinaNet (Lin et al., 2020). These detect objects based on features extracted by a backbone and regions of interest (ROI) generated by dense grid sampling. They are characterised by near real-time detection speeds.

Second, there are two-stage detectors, such as R–CNN (Girshick et al., 2014) or Faster R–CNN (Ren et al., 2017). In contrast to single-stage detectors, ROIs are not generated by fixed cells, but rather, they are generated by a trained region proposal network (RPN). These methods achieve higher accuracy rates but are generally slower than single-stage methods. For segmenting objects, the two-stage detector Faster R–CNN was extended with a segmentation head for generating object masks inside the bounding boxes. The resulting novel method, Mask R–CNN (He et al., 2017), is an often used state-of-the-art instance segmentation. In addition to the techniques already described in this section, novel transformer-based methods, such as the detection transformer (DETR) (Carion et al., 2020) have been developed for detecting and segmenting objects. These take advantage of an enlarged receptive field and, as a result, outperform the two-stage detector Faster R–CNN.

Recently, some of the described deep-learning based object detectors have been applied in studies for single tree detection and segmentation, thereby forming a third category of new segmentation methods. The study by Weinstein et al. (2019) demonstrated individual tree crown detection in the form of bounding boxes using the one-stage detector RetinaNet and aerial imagery data with a resolution of 0.1 m. The research site was located in California (USA) and contained an open woodland forest of live oak (*Quercus agrifolia*) and foothill pine (*Pinus sabiniana*). The labels used for training and validation were generated by either a lidar-based segmentation method from Silva et al. (2016) or/and by manual annotation. Under the premise of an intersection over union (IoU) of at least 50%, they reported an F1 score of 65%. Many of the false-positive tree segments were a result of inconsistencies in the training data between the unsupervised lidar labels and the hand annotation labels. In another study, G. Braga et al. (2020) applied Mask R–CNN to segment individual trees in a tropical forest in Brazil using WorldView-2 satellite imagery data with a resolution of 0.5 m. They achieved an F1 score of 86% on a test plot using reference data generated by visual interpretation. However, over-segmentation is still a problem that tends to increase with the size of the tree crown.

Compared to satellite or aerial data acquisitions, UAV-based missions offer a distinct advantage. The achieved resolutions of the remote sensing data are higher and therefore offer much more accurate detail. However, only small areas can be surveyed using these drone systems (Diez et al., 2021). Chadwick et al. (2020) detected regenerating coniferous trees under deciduous trees in leaf-off condition using UAV-based imagery. The research site, which is located in the Rocky Mountains valley (USA), is dominated by lodgepole pine (*Pinus contorta*), white spruce (*Picea glauca*), and aspen (*Populus tremuloides*). The researchers used RGB images with a resolution of 0.03 m and achieved an F1 score of 91% using Mask R–CNN as the segmentation tool. Their study showed that regenerating coniferous trees can be detected effectively. A method presented by Windrim and Bryson (2020) detected single trees in lidar-based images using Faster R–CNN in two New South Wales (Australia) pine forest stands (*Pinus radiata*). The forest stands are characterised by stem densities of 400 and 600 stems/ha. They achieved detection accuracies with F1 scores of between 76% and 93%. In the denser forest stand, the method was clearly inferior to conventional watershed segmentation and still has problems with under-segmentation.

Applications of transformer-based approaches for single tree detection and segmentation are rare. For instance, the method used by Chen and Shang (2022) counted trees in satellite RGB imagery using a model called density transformer (DENT). The data sets collected from different locations across the United States contain a variety of different tree species, including white oak (*Quercus alba*) and shortleave pine (*Pinus echinata*). When compared with YOLOV3 (Redmon and Farhadi, 2018) and Faster R–CNN, DENT achieved mean absolute error values that were better by 30% and 50%, respectively. The study from Dersch et al. (2022) demonstrated tree detection based on the transformer-based object detection DETR. The research site was a temperate forest in Germany with a tree density of around 230 stems/ha consisting of European beech surrounded by Norway spruce (*Picea abies*). F1 scores were evaluated for a coniferous plot (83%), a mixed plot (86%) and a deciduous plot (71%). In the mixed plot, DETR outperformed by more than 15% when compared to YOLOV4.

In summary, there has been in recent years a focus on deep learning based approaches in tree segmentation research due to the potential increase in accuracy. One of the most frequently investigated methods of instance segmentation is Mask R–CNN. To best of our knowledge, transformer-based instance segmentation has not yet been investigated for tree delineation. Transformer-based approaches are characterized by an enlarged receptive field and a lightweight architecture. Thus, it is crucial to determine whether the characteristic transformer mechanism of self-attention provides an advantage over well-known instance segmentation methods (e.g. Mask R–CNN) in single-tree segmentation, as well as the challenges that arise in its adaption and application. Moreover, mainly multispectral images are used as input data, and they only represent the upper visible forest canopy structure. Therefore, the majority of detected and segmented trees are dominant trees. However, this limitation could in principle be mitigated by applying lidar and thereby using lidar-based metrics such as point density, CHM, lidar intensity and penetration rate. In general, effects such as under-segmentation and over-segmentation are still a problem for deciduous and coniferous forests. Moreover, the quality of the segmented tree polygons, which are potentially better due to using a trained neural network, has not addressed and investigated. For instance, Briechle et al. (2021) noted that precisely delineated tree polygons are essential for improved tree species classification. Most of these studies used only visually generated reference data to test the CNN model, without resorting to field measurements that allow the tree canopy polygon to be digitized as precisely as possible. This is in full accordance with the study from Kattenborn et al. (2021), which reported that 62% of publications between 2017 and 2020 that referred to CNN-based remote sensing methods used reference data labelled by visual interpretation.

The novelty of our study is the application of instance segmentation for tree segmentation on multispectral images and lidar data acquired in the same study area. We used the neural networks Mask R–CNN and

DETR, thereby comparing a conventional and transformer-based instance segmentation along with traditional segmentation methods. Lidar images reflecting the 3D forest structure are combined with multispectral channels (red, green, blue and near infrared). Because we have exact tree crown polygons generated with the help of reference data from field measurements, we can investigate the quality of the segmented tree crowns as well.

2. Materials

2.1. Study area

Our study area was located 35 km northeast of Munich, near the Kranzberg Forest Roof Experiment (KROOF) research site, located at 11°39‘42″ E, 48°25‘12″ N. The mixed forest area, administered by the Bayerische Staatsforsten, is characterised by spruces and large groups of beeches. The stem density in this forest varies in a range of 181–236 trees/ha, with tree heights between 19 m and 36 m.

Field measurements were carried out to obtain reference data for evaluation. The positions of dominant trees that had a minimum breast height diameter (BHD) of 15 cm, were measured by tacheometric means with an accuracy of 2 cm. The BHD was measured using a standard caliper. The coniferous plot (Fig. 1, Plot #1) was dominated by large coniferous trees and partly by some understory trees. The mixed plot (Fig. 1, Plot #2) was characterised by 60% coniferous and 40% deciduous trees. The third plot (Fig. 1, Plot #3) was composed of 76% deciduous and 24% coniferous trees of different sizes and ages and is referred to as deciduous plot in the following due to the high proportion of deciduous trees. The plot characteristics are provided in Table 5.

2.2. Data acquisition and preparation

2.2.1. Aerial multispectral data

We collected multispectral aerial images in August 2020 and July 2021 using a RedEdge MX dual camera system (RedEdge-MX, 2020) attached to a remote-controlled hexacopter (DJI M 600 Pro). Additionally, we mounted an upward-facing light sensor for accurate ambient light calibration. The camera system captured 10 channels (spectral range 475–842 nm) with a focal length of 5.5 mm and a horizontal field of view (HFOV) of 47.2°. In order to achieve a radiometric calibration, images of a calibration panel were taken before the flights. The flight height for the mission in August 2020 was 90 m and the flight height for the mission in July 2021 was 80 m, resulting in ground sample distances (GSDs) of 5.93 cm and 5.30 cm, respectively. The flight speed for both missions was 5 m/s above ground. The software MetaShape (2010) was used for generating true orthophotos (TDOPs). The postprocessing steps included (a) radiometric calibration of the images, (b) bundle adjustment, (c) point cloud generation and (d) orthomosaic generation. We exported four channels (red at 612 nm, green at 560 nm, blue at 475 nm, and near infrared at 842 nm) from the multispectral camera and added an extra PAN channel by weighting the red, green and blue channels with the values 0.2, 0.6, and 0.2, respectively. Finally, a five-channel TDOP with a cell size of 5 cm was exported to be available for further processing. Table 1 provides the details of the photogrammetric campaign.

2.2.2. Lidar data

In addition to the multispectral imagery, we collected lidar data

Table 1
Flight parameters of aerial image acquisition and software packages used.

Multispectral camera	RedEdge MX
SFM - Software	MetaShape (MetaShape, 2010)
Field of View (degree)	47.2
End lap (%)	90
Side lap (%)	60
Acquisition time	August 2020/July 2021
Images	3770/3560
Flight height (m)	90/80
GSD (cm)	5.9/5.3

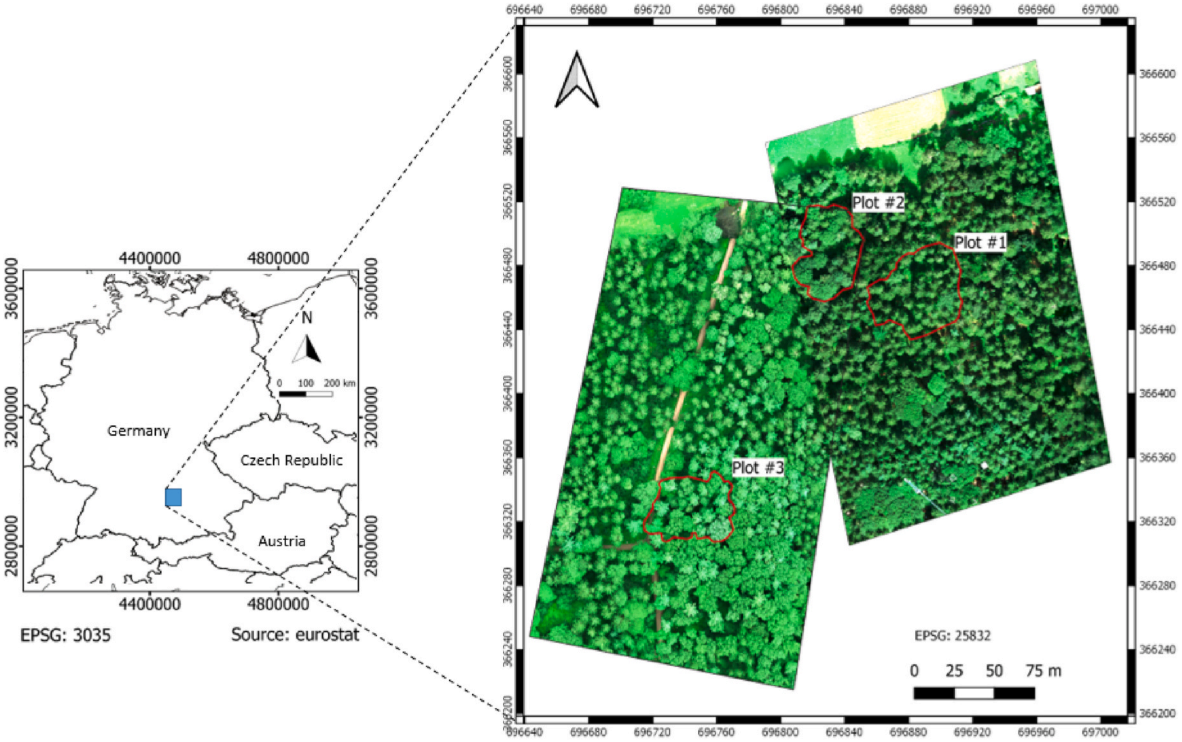


Fig. 1. RGB true orthophoto (TDOP) showing the research area of KROOF containing three plots: #1, #2 and #3. The remaining forest area was used for training and validation.

using a RIEGL miniVUX-1UAV (RIEGL, 2020) instrument, which was also mounted to the drone platform. For each day of flight in August 2020 and July 2021, three flights were conducted at a 90 m flight height with a lidar strip distance of 22 m, resulting in a side lap of more than 50%. The repetition frequency of 25 kHz added up to an average point density of 560 pt/m^2. For accuracy reasons, a stationary global navigation satellite system (GNSS) receiver was used to provide correction data for the subsequent differential GNSS (DGNSS) data processing. Moreover, a calibration flight was performed over a large property with several buildings to collect data to verify the boresight parameters. Furthermore, the buildings' enclosing polygons were measured by DGNSS for the exact transformation of the lidar data into the reference coordinate system (=UTM 32N = EPSG 25832). Finally, we aligned the lidar strip, taking into account misalignments due to remaining errors in the boresight calibration and INS-based drift effects. The processing steps to produce the final lidar point cloud were as follows: (a) Boresight calibration (b) Generation of lidar data for each mission in the reference coordinate system (c) Lidar strip alignment (d) Merging of the lidar strips (e) 3D correction of complete lidar data set using reference polygons (f) Calibration of lidar intensity. Moreover, the bare ground of the forest (DTM) was filtered out of the lidar point cloud using Terrasolid software (TerraSolid, 2020). In the last step, we generated a surface model from lidar points (DSM_{lidar}) at a grid spacing of 5 cm. For further processing, we discarded all lidar points within a height bound of 10 m above ground to avoid the impact of bushes and shrubs. Table 2 summarizes all of the details of the lidar campaign with reference to the software packages used.

2.2.3. *Calibration of lidar intensity*

The full waveform scanner provided waveforms that were decomposed using GeoCode software (LasTools, 2021), resulting in information regarding the reflected intensity. Thus, the 3D coordinates (x_P, y_P, z_P) of each reflecting object P hit by the laser pulse were obtained in combination with the intensity I_P. Overall, this decomposition generated a point cloud for the forest area represented by the vector $X_n(x_n, y_n, z_n, I_n)$, $n = 1, \ldots, N$ (N is the total number of points in the point cloud). The intensity I_n depends on the traveling distance r_n (in one direction) and can be calibrated using a data-driven model (Höfle and Pfeifer, 2007).

$$I_n^{corr} = \frac{I_n}{1+(r_n-r_0)C1+(r_n^2-r_0^2)C2} \tag{1}$$

In order to obtain reasonable values for the parameters $C1$ and $C2$, samples for I_n (see Eq. (1)) were taken at two flying heights, r_1 and r_2, as the mean value from three small concrete areas (ca. 4 × 4 m^2) located in the vicinity of the buildings (see Figure A1 in the Appendix). The value r_0 describes the reference distance of 50 m. Note that the intensity correction provided the final lidar point cloud for the entire test site that was subsequently used for (1) generating lidar-based images (see Subsection 2.3), and (2) calculating tree segments using four baseline methods (see Subsection 4.1). Fig. 2 shows the effect of the intensity calibration.

Table 2
Flight parameters of lidar flight and software packages used.

Acquisition time	August 2020/July 2021
Scanner type	RIEGL miniVUX-1UAV (RIEGL, 2020)
Platform	DJI M 600 Pro
Spectral wavelength (*nm*)	1550
Pulse repetition frequency (*kHz*)	100
Beam divergence (*mrad*)	0.5
Flight speed (*m/sec*)	5
Flight height (*m*)	90
Side lap (%)	60
Point density (pts/m^2)	ca. 560
Footprint size (*mm*)	45
Area (ha)	6
Software for strip alignment	StripAlign (BayesMap, 2018)
Software for GNSS/INS	Inertial Explorer (Novatel, 2018))
Software for lidar georeferencing	GeoCode (LasTools, 2021))
Software for DTM filtering	Terra Scan (TerraSolid, 2020))

2.3. *Data fusion*

The two instance segmentation methods in this study require images as input data. Therefore, in addition to the images from the multispectral camera, the lidar point cloud based layers were projected into 2D bird's-eye views with a resolution of 5 cm × 5 cm and then all stored together in a data cube. First, we generated a canopy model from lidar (CHM_{lidar}) and a photogrammetric canopy model (CHM_{photo}) from the DSM_{lidar} and photogrammetric surface model (DSM_{photo}), each being normalized with the DTM. We also generated a data layer point density (P_DENSE) with a footprint of 5 cm × 5 cm. Each pixel value represents the number of lidar points within the voxel above the footprint from the bottom to the top of the CHM_{lidar}, thereby describing the penetration of the laser beam in the vegetation. In addition, we processed a data layer mean intensity (M_INTEN) representing the mean lidar intensity in a voxel of the same dimensions.For our experiments, we defined four channel combinations (Table 3). The idea was to generate two optical combinations (*RGB*, *CIR*), one lidar-based combination (*LIDAR*) and one combination using the optical PAN channel that was converted from channels red (612 nm), green (560 nm) and blue (475 nm) and two lidar channels ($P + LIDAR$). Note that all channels were enhanced using histogram equalization.

2.4. *Evaluation data*

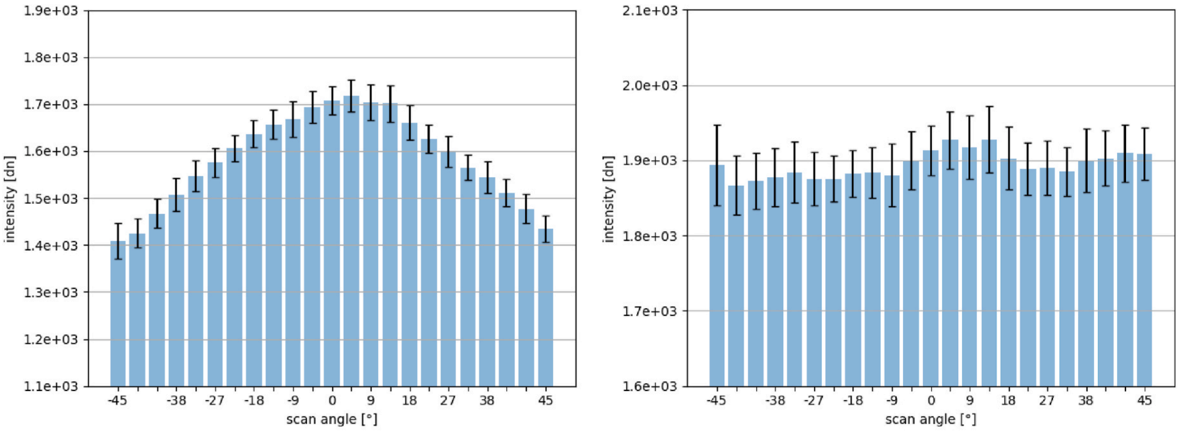
The research area was split into three parts for training, validation and testing of the networks. Precise tree segments for testing were selected in plots #1, #2 and #3, considering the field measurements. Training and validation data are required to optimize the instance segmentation methods. Therefore, the remaining area was labelled by visual interpretation and used for training and validation. Table 4 shows the tree parameters for training and validation.

2.4.1. *Field survey*

The goal of the field survey was to measure tree positions as precisely as possible in order to generate accurate testing data. Due to the expected shading effects in dense forest areas using GNSS systems, a survey campaign was conducted in April 2021. First, a traverse was measured in the area of the three plots #1, #2 and #3. The traverse included seven polygon stations and was georeferenced using three geodetic points. The instruments used included the Trimble R12i GNSS system and the Leica TCRP1203+ total station. Afterwards, tree positions were surveyed from the polygon points by tacheometric means. The BHD of each dominant tree greater than 10 cm was also measured using a scale bar, and the tree group was also documented. In summary, we surveyed 55 trees in plot #1, 36 trees in plot #2 and 34 trees in plot #3 (see also Table 5). The estimated accuracy of the tree positions was less than 10 cm.

2.4.2. *Labeling of tree crowns*

The reference data are provided in the form of individual tree segments as enclosing polygons. For this purpose, the 3D point cloud and the RGB TDOP were used for visualization. For labelling the training data, tree segments were first defined in the orthophoto and refined or enlarged because of shadows in the RGB TDOP with the help of the CHM_{lidar}. This ensured to measure the true outermost crown shape, thereby providing a good approximation of the real crown radius. In a few cases, the 3D point cloud was also used to separate and delineate trees that stood close together because they could not be clearly distinguished.The tree positions in the 3D point cloud were visualised to provide the test data in oblique view. Tree polygons were then digitized in the 3D point cloud and superimposed on the RGB TDOP. This linked the precise tree position from the field measurements to an accurately measured tree segment and guaranteed high-quality reference data for



(a) (b)

Fig. 2. Lidar intensities in a profile across a lidar strip captured at a flying height of 60 m. Distance parameters were: $r_0 = 50.0$ m, $r_1 = 50.0$ m, $r_2 = 90.0$ m. (a) Before calibration. (b) After calibration.

Table 3
Definition of channel combination used in the experiments. Numbers in brackets are in nm and indicate wavelength.

Definition	Channel 1	Channel 2	Channel 3
RGB	Red (612)	Green (560)	Blue (475)
CIR	NIR (842)	Red (612)	Green (560)
LIDAR	CHM_{lidar}	*P_DENSE*	*M_INTEN*
P + LIDAR	PAN	CHM_{lidar}	*P_DENSE*

Table 4
Tree parameters of areas for training and validation.

Parameter	Training	Validation
Size (m^2)	52635	12364
Trees	1185	223
Trees/ha	225	181
Forest type	mixed	mixed
Tree heights (m)	19–36	20–36

Table 5
Tree parameters of reference plots for testing.

Parameter	plot #1	plot #2	plot #3
Size (m^2)	2434	1883	1840
Trees	55	36	34
Trees/ha	226	191	185
Forest type	coniferous	mixed	mixed
Tree heights (m)	19–35	20–34	19–34

verifying the accuracy of the segmentation method. Fig. 3 shows an example of labelled trees in the lidar point cloud.

3. Methods

3.1. Outline of methods

The basic idea of our approach is the delineation of single trees by instance segmentation from multispectral images and images generated from lidar data. Fig. 4 shows the complete schematic workflow. Multispectral imagery and lidar data were acquired from a low-flying UAV with high point density. We combined multispectral imagery and images from lidar data into a data cube that stores images separately into channels with the same resolution. Three channels were selected from the data cube as input data. As instance segmentation methods, we used Mask R–CNN and DETR with ResNet-50 (He et al., 2015) as a backbone pre-trained on the ImageNet data set (Deng et al., 2009). The best model was found using the training and validation data set. Reference segments were generated by labelling tree crowns using the lidar point cloud, the RGB TDOP and tree positions acquired in the test site by tacheometric means. Evaluation metrics such as accuracy, F1 score, recall and precision were calculated using both reference segments, true tree positions from field measurements and the detected segments. Furthermore, we calculated the segmentation quality using the mean IoU between true positive and reference segments.

3.2. Mask R–CNN

The instance segmentation Mask R–CNN used in this study is based on the two-stage object detection Faster R–CNN Ren et al. (2017), which is the second improvement of the original R–CNN Girshick et al. (2014) object detection. Unlike the two previous versions, R–CNN and Fast R–CNN, Faster R–CNN uses an RPN. Mask R–CNN is based on the Faster R–CNN method and has been modified to include a masking head. Each detected object is additionally delineated, and thus, a mask is determined. In combination with Faster R–CNN, this instance segmentation represents a fully deep-learning based end-to-end trainable pipeline. More detailed, the pipeline Mask R–CNN (see Fig. 5) consists of two steps. First, region proposals were generated using the backbone CNN and the RPN. ResNet-50 was used as backbone and the different stages are referred to as c2 through c5, with the resulting feature maps labelled p2 through p6. This RPN is a trainable, independent object detector that performs class-independent object detection. It is important to mention that the CNN features provided by the backbone were used for the RPN and the rest of the object classifications. In the second stage, the feature maps of the detected region proposals were extracted using the region of interest (ROI) alignment in a fixed size. As the last step, the classes, bounding boxes and masks of the detected objects were determined. The tasks were performed in separate heads. As a result, all target objects in the scene were exported as masks and class labels.

3.3. DETR

DETR processes global image information using the transformer mechanism (Vaswani et al., 2017). This approach considers object detection and segmentation as a direct set prediction problem. Moreover, it eliminates several sub-tasks that require prior knowledge about the problem, such as anchor generation. The global image context and the relationship between objects are crucial factors. Predictions can be determined in parallel and directly using a small set of learned object

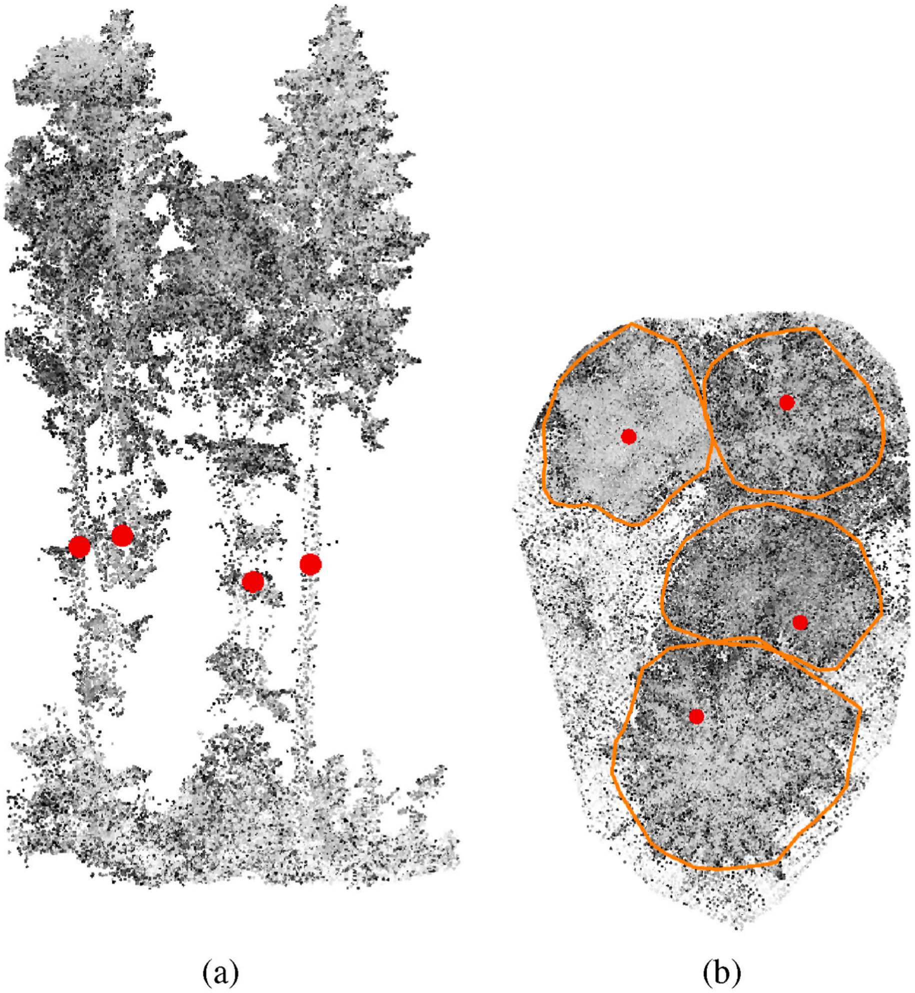


Fig. 3. Labeling of four single coniferous trees in 3D point cloud for test data. Red dots indicate the positions of corresponding reference trees. (a) Side view. (b) Top view. The manually digitized segments are circled in orange. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

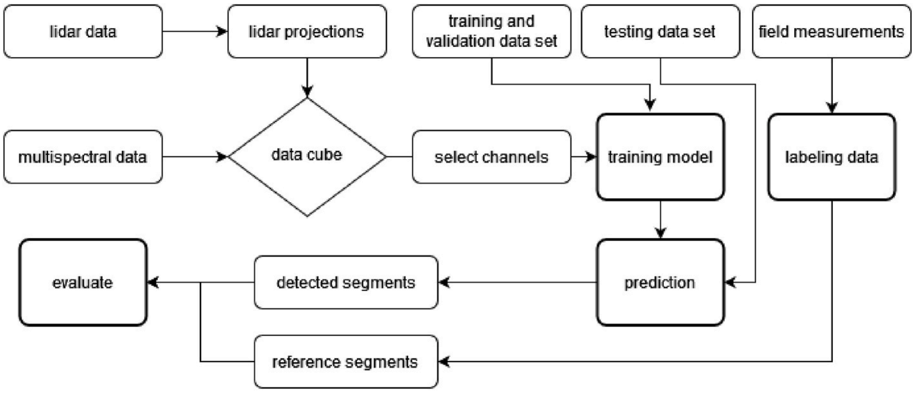


Fig. 4. The complete process of segmenting single trees using Mask R–CNN and DETR.

queries. Fig. 6 shows the main elements of the overall DETR architecture. First, ResNet-50 provided features and positional encoding was added to the CNN feature map. Finally, the features were transferred to a transformer encoder, followed by a transformer decoder that generated N object queries. Lastly, the bounding boxes were predicted and the associated classes determined using a feed-forward network (Carion et al., 2020). Finding and evaluating ground truth and predicted boxes is a vitally important task. Bipartite matching, which was solved using the Hungarian algorithm (Kuhn, 1955), defined the set prediction. Here, ground truth boxes and a larger set of predicted boxes were matched efficiently. The loss for matched pairs, called the Hungarian loss, is a linear combination of a negative log-likelihood for class prediction, the generalized IoU loss (Rezatofighi et al., 2019) and the commonly used l_1 loss.DETR can be extended to perform instance segmentation. The mask head can either be trained in parallel with object detection as a one-step process, or in a two-step approach, where an object detection model is

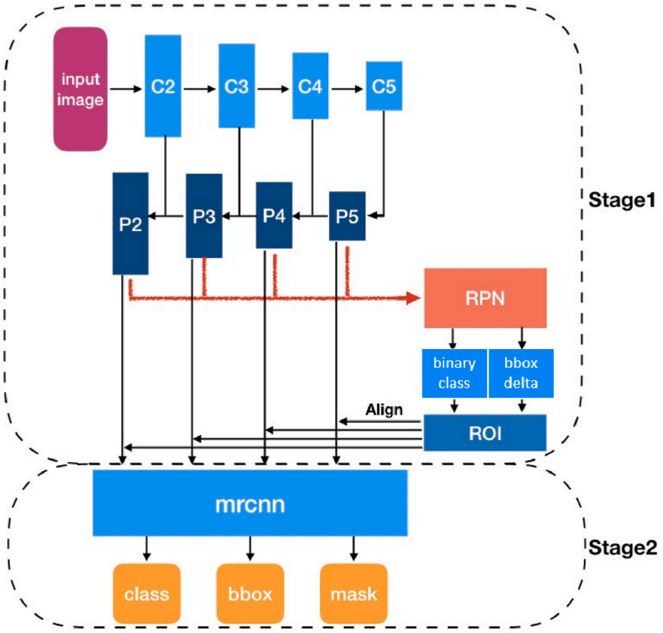


Fig. 5. Schematic overview of Mask R–CNN. Image from Zhang (2022).

trained first, and afterwards, a mask head is trained to discriminate between instances and backgrounds within bounding boxes. Subsequently, all weights are frozen and only the mask head is trained for approximately 25 epochs. Studies yield similar results for these two approaches. In this study, we trained the object detection head and mask head in a one-step process. Therefore, two pre-trained models were loaded, one for object detection and one for panoptic segmentation. For parallel training, the AdamW optimizer was modified to implement a separate learning rate for the segmentation. Thus, both models were trained and optimized with optimal learning rates.

3.4. Choice of parameters

For instance segmentation Mask R–CNN, we used the Python implementation of detectron2 (Wu et al., 2019). We adapted the parameters learning rate to 1e-4, the batch size to 12 and the maximum number of epochs to 160.For instance segmentation DETR, we used the GitHub repository of Carion et al. (2020) as codebase. Most of the hyperparameters correspond to the default configuration. In this study, we modified the backbone learning rate, the learning rate of the encoder/decoder and the panoptic learning rate to 4e-7, 5e-6 and 4e-6, respectively. We changed the batch size to 2 and the decay rate to 1e-4. We selected the best model within 80 epochs, considering overfitting effects. The models of both instance segmentation methods were trained and validated using the data set described in 2.4.2. Random data augmentation (horizontal/vertical flip, rotation, brightness/contrast/saturation change, image crop) was applied to improve the training and, therefore, the model. Both instance segmentation methods achieved best results using an unfrozen ResNet-50 backbone. Attempts to improve the result using a training strategy using frozen backbone layers showed no improvements.

4. Experiments

4.1. Experimental setup

The experiments were divided into several sections. Experiment #1 focused on the performance of delineating trees by Mask R–CNN and DETR. The four channel combinations defined in Section 2.3 (see Table 3) were used. The entire test site's data set was subdivided into training, validation and testing, and three areas were selected for testing. The remaining data set was subdivided into 80% training and 20% validation. The total imagery was tiled into 512 × 512 pixel size images with 50% overlap. Experiment #2 focused on comparing the best results of experiment #1 with four baseline methods, normalized cut (*NCut*) (see Reitberger et al. (2009)), *Silva* (see Silva et al. (2016)), *Li* (see Li et al. (2012)) and watershed segmentation (*WS*) (see Roussel and Auty (2022)) with area-based evaluation applied (see Section 4.2). Moreover, the quality of the segmentation was examined in the third experiment. Note that the methods *NCut* and *Li* used the entire lidar point cloud as input data, while *Silva* and *WS* used the data set CHM_{lidar} generated from the lidar point cloud.We applied the lidar package lidR (Roussel et al., 2020) for methods *Silva*, *Li* and *WS*. The TreeFinder software package (PrimaVision, 2022) was used for method *NCut*. An Ubuntu workstation equipped with 256 GB of RAM, an Nvidia RTX 8000 GPU and an AMD Ryzen Threadripper 3970X processor was used for processing the data.

4.2. Evaluation and accuracy assessment

The evaluation of the experiments was performed using an area-based method that uses the IoU of reference and detected tree segments as a basis. The detected tree is counted as true positive (*TP*) if the IoU value is above a threshold of 50%. The detected tree segments that could not be assigned to a reference segment are counted as false positives (*FP*). It should be noted that tree heights are not considered for evaluation in this work. Furthermore, reference trees are marked as false negatives (*FN*) if no corresponding reference tree is found. Finally, the metrics of accuracy, recall, precision, F1 score and IoU were calculated for each test plot.

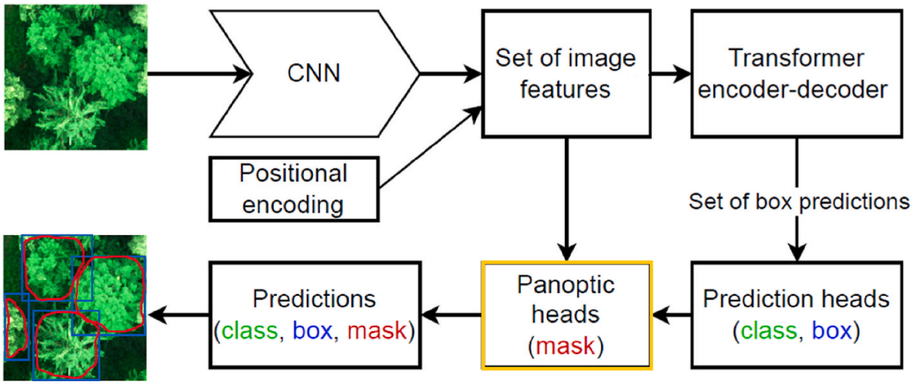

CNN
Set of image features
Transformer encoder-decoder
Positional encoding
Set of box predictions
Predictions (class, box, mask)
Panoptic heads (mask)
Prediction heads (class, box)

Fig. 6. Schematic overview of the instance segmentation method DETR. Orange box indicates the masking heads. Adapted from the original publication (Carion et al., 2020). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

$$accuracy = \frac{TP}{TP + FP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{5}$$

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

Finally, in order to show the quality of the segmentation, the mean IoU ($mean_{IoU}$) was calculated for each plot using the sum of the IoUs over all *TPs* divided by the number *n* of all *TPs*:

$$mean_{IoU} = \frac{1}{n} \sum_{i=1}^{n} IoU_i \tag{7}$$

4.3. Results

4.3.1. Instance segmentation using mask R–CNN and DETR

In experiment #1, we demonstrated the performance of the instance segmentation approaches Mask R–CNN and DETR in Fig. 7. The plots in the left column (7a, 7d, 7g), middle column (7b, 7e, 7h) and right column (7c, 7f, 7i) show the results of Mask R–CNN, DETR and four baseline methods, respectively. The plots in the first row (7a, 7b, 7c), second row (7d, 7e, 7f), and third row (7g, 7h, 7i) show the results of test plot #1 (coniferous), plot #2 (mixed), and plot #3 (deciduous), respectively.

First, we investigated the image-based channel combinations *RGB* and *CIR* using Mask R–CNN (see Fig. 7a, d, 7g). The best channel combination, *CIR*, achieved an F1 score of 92% in the coniferous plot, 83% in the mixed plot and 85% in the deciduous plot. The precisions in the mixed and deciduous plots were about 8%–9% worse than the coniferous plot. In the coniferous plot, using the channel combination *RGB*, we obtained an F1 score of 92% and an accuracy of 85%. Recall and precision also achieved values of 92% and 93%, respectively. Compared to the coniferous plot, the mixed and deciduous plots contained a large number of deciduous trees. The F1 scores of these plots were 8%–9% worse than in the coniferous plot with 83% and 84%, respectively. We obtained the worst precision in the deciduous plot with 82%. In summary, we observed that the channel combination *CIR* achieved the best results.

Comparing the results of Mask R–CNN obtained with the optical imagery, we now focused on the outcomes of Mask R–CNN using the lidar-based channel combinations *LIDAR* and *P* + *LIDAR* (See two right-hand bar combinations in Fig. 7a, d and 7c). The channel combination LIDAR, consisting of three solely lidar-based channels, achieved F1 scores that were similar to the *CIR* data set in the coniferous forest and mixed plot with minor differences of 2%. However, the results differed significantly in the deciduous area by about 4%. Finally, we considered the channel combination *P* + *Lidar*, which uses an optical channel combined with two lidar channels (see 3). We found that in the coniferous and mixed plots, the results were 2% worse for both plots compared to the *CIR* channel combination. However, the significant drop in F1 score for this hybrid channel combination is remarkable, with 10% in the deciduous plot.

The results obtained by the instance segmentation DETR, summarizing Fig. 7a and b, were similar to Mask R–CNN in the coniferous plot. In the channel combinations *RGB* and *CIR*, DETR was about 2% worse in F1 score. The results of DETR with the channel combinations *LIDAR* and *P* + *LIDAR* are around 2%–3% worse. In the mixed plot (Fig. 7f and e), DETR performed 6% and 1% better using the optical channel combinations *RGB* and *CIR* than the F1 scores of Mask R–CNN. We also noted similar results for the channel combinations *LIDAR* and *P* + *LIDAR*. The results for the deciduous plot (Fig. 7g and h) showed a significant deterioration in F1 score using DETR with the channel combinations *RGB* and *CIR* of about 8% and 4%, respectively. The channel combinations *LIDAR* and *P* + *LIDAR* also led to an accuracy decrease of about 5%. However, the lidar-based data set results did not indicate any significant differences.

4.3.2. Comparison to existing baseline methods

In this section, we focus on comparing the two instance segmentation methods with four selected baseline methods (see Fig. 7c, f and 7i). Note that these methods only use the lidar point cloud as input data. Because the baseline methods depend on control parameters, we optimized the most important ones in a sensitivity analysis by coarsely varying the parameter values within a reasonable range and maximizing the F1 score in a grid search (see Table A2 in the Appendix).

The results of the four baseline methods showed that they were outperformed in all test plots by Mask R–CNN and DETR. The four baseline methods had their best results in the coniferous test plot. However, when compared to Mask R–CNN, *NCut* scored 13% worse, *Silva* scored 38% worse, *Li* scored 76% worse and *WS* scored 58% worse in the coniferous plot. In the mixed plot, all of the baseline methods deteriorated, particularly *NCut* by 26%, but also *Silva* by 6%, *Li* by 2% and *WS* by 17%. As expected, the effect of accuracy deterioration in the deciduous plot was even more evident for *NCut* with 16%, *Silva* with 1% and *WS* with 5%. However, *Li* achieved a 4% better result.

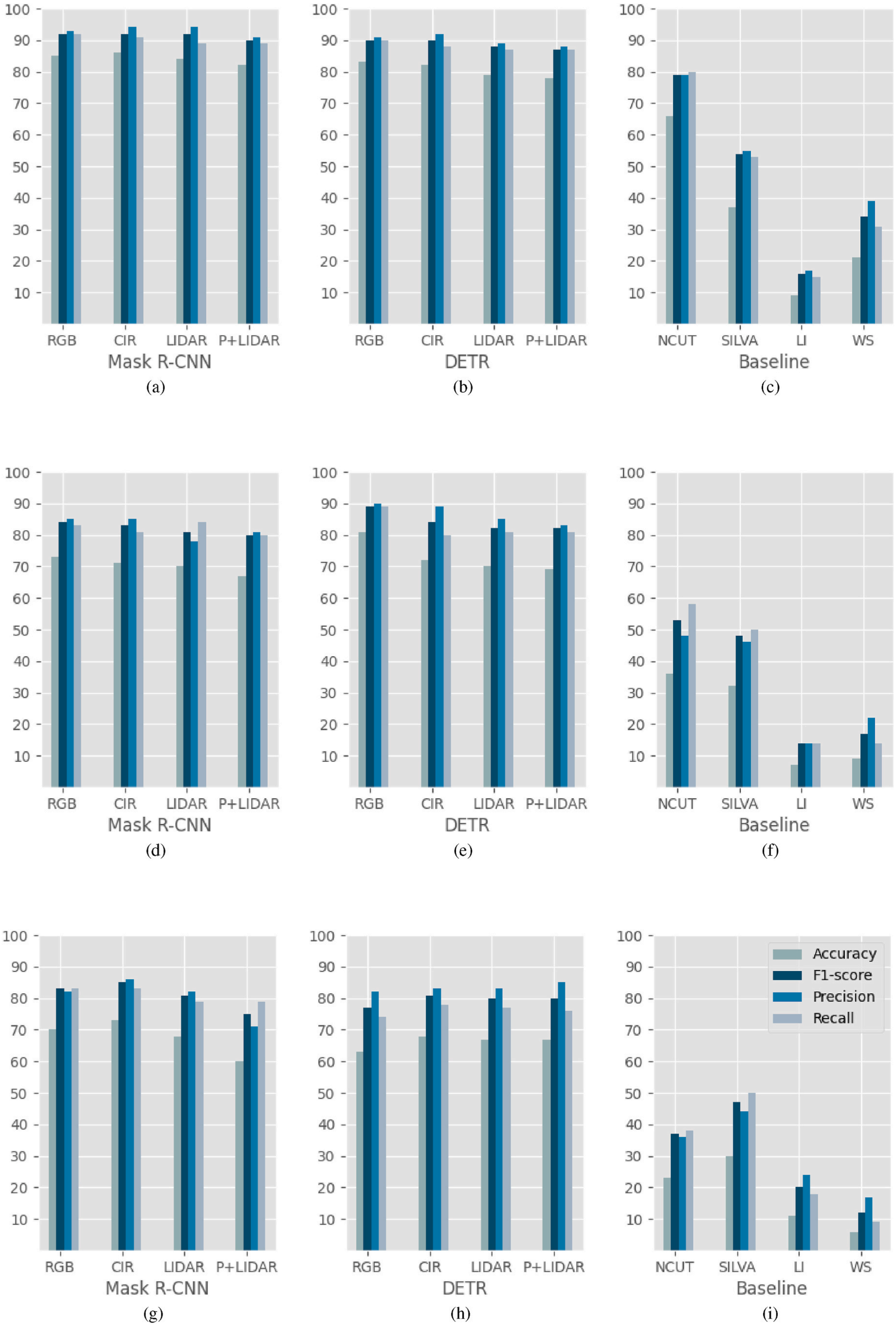
4.3.3. Quality of segmentation

In the last experiment, #3, we investigated the quality of the segmentation results using the $mean_{IoU}$ (see Eq. (7)). For the instance segmentations Mask R–CNN and DETR, we limited the analysis to the best performing channel combination *CIR*. As with the experiments discussed in Section 4.3.2, we selected *NCut* as the representative method. Table 6 shows clearly better values for the instance segmentations Mask R–CNN and DETR. More precisely, the two instance segmentations performed 10%–16% better than the baseline method *NCut*. In addition, the number of true positives was higher for the instance segmentations and lower for *NCut*. Fig. 8 gives examples of the quality of single tree delineation in a mixed forest area.

5. Discussion

5.1. Comparison of the instance segmentation results

The instance segmentations had only minor differences between the multispectral channel combinations *CIR* and *RGB* and the different forest types. Interestingly, the results of the lidar-based combinations tended to be worse for Mask R–CNN and DETR in the coniferous and mixed plots. However, in the deciduous plot, the results for DETR were better than the visual ones for the lidar-based channel combinations. In summary, we found no significant differences among either the instance segmentations or the optical and lidar-based channel combinations. In other words, the lidar-based information in the form of depth images (*CHM_{lidar}*), lidar intensity (*M_INTEN*) and lidar point density (*P_DENSE*) did not seem to present any added value. Also, the combinations of depth images, lidar intensity and PAN channel did not provide any improvement. In principle, the point density provides the stem information to the neural network in case of high lidar point density. We suspect that the number of labelled tree stems was insufficient to adequately train the highly parameterized network concerning the stem information. To overcome this drawback, we could train two backbones with multispectral and lidar-based images in parallel and use a feature vector merged from both backbones. Alternatively, significantly more training data could lead to a more optimized model that better represents the lower forest structures, including the tree stems.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 7. Results of instance segmentation with Mask R–CNN in the left column, DETR in the middle column and baseline methods in the right column. Rows one, two and three show the results of test plots #1 (coniferous), plot #2 (mixed) and plot #3 (deciduous), respectively. Y-axis is percentage.

Table 6
Quality of segmentation (see Equation (7)) for instance segmentation methods Mask R–CNN and DETR (using the data set *CIR*) and *NCut* (using the lidar point cloud) as mean value for each plot. Values in brackets indicate numbers of true positives. Numbers are in percent.

	Coniferous	Mixed	Deciduous
Mask R–CNN (*CIR*)	77 (50)	78 (29)	75 (28)
DETR (*CIR*)	77 (48)	76 (29)	77 (27)
NCUT	69 (44)	62 (21)	65 (13)

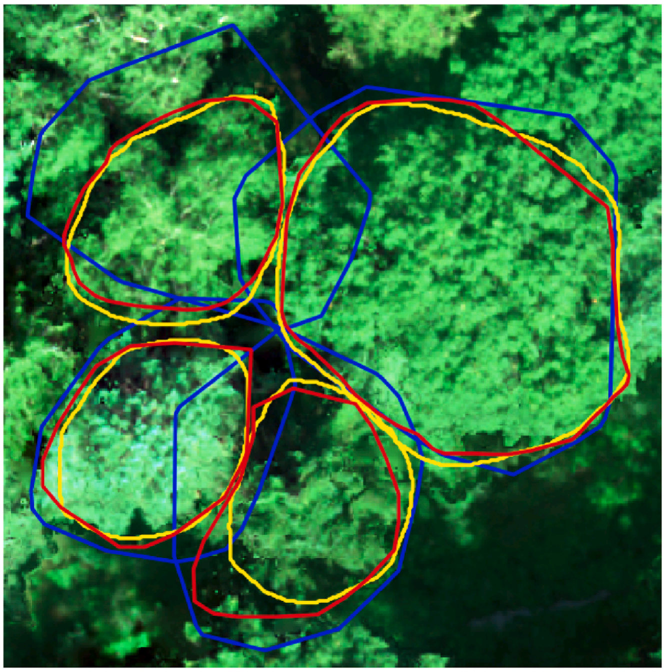


Fig. 8. Examples of the single tree delineation by Mask R–CNN in yellow, *NCut* in blue and reference tree segments in red. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

We expected that DETR would outperform Mask R–CNN. However, the methods had nearly identical performances. We identified the reason for this was that the number of training samples was too small to sufficiently train the implementation of DETR we used. Moreover, one of the most important parameter *object queries* optimized for the COCO data set could not be changed. The authors of DETR reported improvements in detection accuracy, especially for large objects. This can be attributed to the processing of global information by self-attention. However, we could not detect any significant improvements in our investigations. In our opinion, this advantage would apply to significantly large trees. DETR has a performance loss in detecting small objects because of the positional encoding of image feature maps that refer to tiles of fixed size. Therefore, the ability to detect smaller objects is reliant on the subdivision of the image feature maps. Moreover, in our investigations, the parallelization of the training process for object detection and panoptic segmentation based on two different models was difficult to adjust. Because the segmentation model was overfitted faster than the object detection model, we had to apply a separate learning rate and match it to the training process of the object detection model. However, the DETR architecture is relatively simple and is less complex than Mask R–CNN and can therefore be adapted effectively. The geometrical characteristics of the detected tree polygon edges seemed a bit smoother using Mask R–CNN compared to DETR. However, the segmentation results of both methods on coniferous and deciduous tree stands were promising, even for nearby trees. An example is shown in Fig. 9 for coniferous stands in 9a and deciduous stands in 9b.

In detail, we noticed in the coniferous plot that Mask R–CNN and DETR detected three out of seven small trees (sizes smaller than 15 m^2). These smaller coniferous trees are often in close range to larger trees, making them difficult to distinguish. Mask R–CNN detected all three big trees (sizes larger than 60 m^2) within the coniferous plot in comparison to DETR, which failed to detect one tree. This was caused by a too-low confidence score causing the tree not to be recognized. In the deciduous plot, we also noticed problems with small trees (sizes smaller than 15 m^2). Out of four small trees, Mask R–CNN recognized one, while DETR failed to recognize any of the four trees. There were also issues with larger trees (sizes bigger than 60 m^2). All six trees in the deciduous plot were detected by both methods, but minor artifacts were generated in addition to the correctly detected trees. These over-segmentation effects are most present for DETR, with four segments, and less so for Mask R–CNN with two segments. In the mixed plot, we observed findings that were similar to the coniferous and deciduous plots. Furthermore, there were occasional problems in the form of under-segmentation in the coniferous and deciduous plots with two trees detected as one. This effect was most pronounced in the mixed plot, with DETR in two cases and Mask R–CNN in three cases. We think these adverse effects could be reduced with more training data and more diverse tree samples.

5.2. Comparison of tree segmentation to baseline methods

The four baseline methods applied to the lidar point cloud performed significantly worse than the instance segmentations for all three forest types. As expected, the results tended to be best for conifers and worst for deciduous trees, as the crowns of deciduous trees merge into a closed structure with no clear maximum.

The baseline methods were unsupervised learning algorithms controlled by parameters determined by the sensitivity analysis. The *Li* and *NCut* methods were used to compute single tree segments based on the entire point cloud and model the tree canopy according to specific tree parameters that control the tree crown shape horizontally and vertically. Both *WS* and *Silva* methods were used to estimate single tree segments based on a filtered CHM; thus, they depended on a smoothing factor. Therefore, these segmentation methods are limited to a few control parameters that describe the variety of tree shapes globally but not individually based on sensitivity analysis. However, the instance segmentation methods were supervised representation learning methods that use the characteristic tree crown shapes of the forest area contained in the training data set. These neural networks do not depend on tree-describing control parameters but on training data prepared by

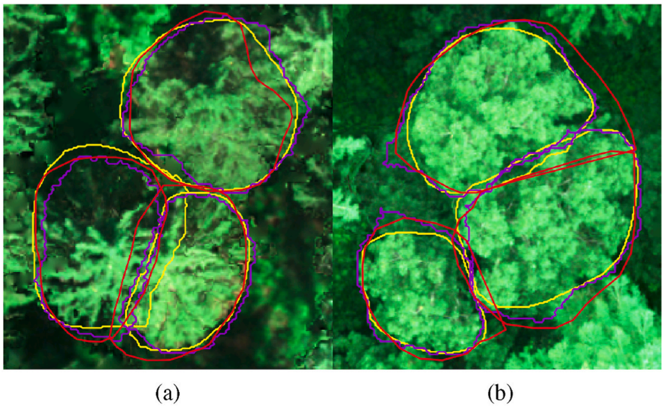


Fig. 9. Example of segmentation results of Mask R–CNN in yellow, DETR in purple and reference in red. Coniferous stands example on the left (a) and deciduous stands example on the right (b). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

experts. This leads to an accurate delineation of the tree canopy shape.

Furthermore, the IoU-based evaluation method, which uses tree crown polygons, also has a decisive influence on the achieved segmentation accuracies. The area-based evaluation method is more rigorous than an evaluation method based purely on tree positions, because it incorporates the tree crown shape of the detected tree. Hence, an IoU of at least 50% is used as a selection criterion for true positives. Some tree segments did not reach the defined minimum threshold of IoU 50% required for evaluation as true positives. However, the mean position (i.e. the estimated tree stem position) seems to have been partially correct.

5.3. *Segmentation quality*

In this section, we discuss the quality of tree delineation using instance segmentation. The instance segmentations Mask R–CNN and DETR delineated single trees much more accurately up to 16% given that the baseline methods are much worse at delineating tree crowns in the surface model. Here, the advantage of neural networks clearly comes into play, as they learn the local forest structure via the labelled training data. Obviously, cast shadows did not play a significant role in the optical images used in this study. In contrast, the baseline methods were based on control parameters that can only be adapted to the respective forest structure after an even more elaborate sensitivity analysis. It should be noted that the enclosing polygons of the trees were measured by visual interpretation of the orthophotos and the 3D point clouds. We assume that the accuracy of this measurement method can be compared to the conventional method of tree crown mapping using an upwards looking mirror system.

5.4. *Comparison to related work*

First, we compared the results of our work with two recent studies using the area-based evaluation method (IoU bigger than 50%). Windrim and Bryson (2020) attained Faster R–CNN Ren et al. (2017) in two pine forest stands with stem densities of 400 stems/ha and 600 stems/ha F1 scores of 93%–76%, respectively. The data set for the present study consisted of lidar-based images, such as vertical density, CHM and average return. For the coniferous plot with a lower stem density of 400 stems/ha, the resulting 93% F1 score was approximately the same as our result of 92% for a stem density of 226 stems/ha. Note that this study utilized object detection based on Faster R–CNN (i.e. bounding boxes).

The study by Hao et al. (2021) applied Mask R–CNN to detect individual trees in a tree plantation using UAV-based imagery. Here, different multispectral channel combinations with at least one channel containing height information were investigated (specifically, red, green, blue wavelengths and digital surface model). The GSD of 0.3 cm was smaller, as in our study. The F1 score results for the total six channel combinations were between 72% and 85%. We achieved slightly higher values in the coniferous plot. It should be noted that the trees of that study's test site grow in a tree plantation with clearly different forest structures than our study site.

Second, we reviewed the study by Krzystek et al. (2020) that reported on experiments conducted at the research site in the Bavarian Forest National Park. In this mixed temperate forest, the percentage of conifers was 43% and of deciduous trees was 57%, with an average stem density of approx. 550 stem/ha. The tree segmentation method normalized cut from Reitberger et al. (2009) was applied to the test area using parameters determined by separate sensitivity analyses for deciduous and coniferous stands. F1 scores for coniferous plots reached 87% and were slightly worse in deciduous plots at 78%. Comparing these results with our experiments, we note that the two instance segmentation methods attained F1 scores around 91% in the coniferous plot and around 83% in the deciduous plot. Clearly, these numbers are significantly better than the results of the study from Krzystek et al. (2020). The evaluation in this study was based on a distance threshold between reference trees and segmented trees. Trees with the smallest distance between the center of a detected tree and a reference tree position were matched. If we applied this point-based evaluation procedure to our test plots, the relevant numbers change to approximately 92% and 72%. The still evident differences between our study and the study by Krzystek et al. (2020) may be attributed to different stem densities and tree species distributions.

Next, we want to address the efficiency and completeness of the current study. The approach required a significant number of manually labelled training samples, although transfer learning was used. In our experience, a TDOP is sufficient for coniferous areas. The labelling in deciduous areas was possible without a 3D laser point cloud, but in some cases, 3D information about the forest structure can help separate closely spaced tree groups. Obviously, field measurements support the complete process of labelling and accuracy assessment. Concerning control parameters, Mask R–CNN and DETR require three primary hyperparameters (i.e. learning rate, batch size and number of training epochs) to be optimized in the learning process. Our experiments required around 9000 iterations for Mask R–CNN and 19,000 iterations for DETR using the workstation described in Section 4.1.

Finally, we would like to address the limitations of the methodology. Both instance segmentation methods only detect trees that are visible in the TDOP, thereby missing regeneration and smaller trees standing in the lower forest layers. Potential solutions could be based on new 3D instance segmentation methods that are based only on airborne lidar point clouds if sufficient point density is available in the lower forest layers. However, these methods still need to be developed. In addition, the lidar-based *Pdense* layer primarily only represents the tree positions of a few coniferous trees. In contrast, no laser point returns can be found on deciduous trees, because their dense tree canopy prevents the penetration of laser beams (see Figure A3 in the Appendix). Thus, the stem information in our data set is insufficient to adequately train the large number of weights of the neural network with regard to tree stems. Second, it turns out that the training process needs a considerably large area of forest to sufficiently train the whole network. We assume that 1000–1500 trees of different species, composition and sizes seem to satisfy the requirements to train and validate the network. For larger forest inventories, this seems to be acceptable. However, for small forest areas captured by a drone in a 15 min flight, the effort exceeds the benefit. A remedy would be more general network models that can be applied to various forest areas with different forest structures. Note that our study does not address the transferability of the neural network to other forest structures.

6. Conclusions and outlook

We presented a study examining the potential of a novel transformer-based instance segmentation approach DETR for single-tree segmentation in a mixed forest area. For the first time, we successfully showed how this new type of network could be adapted and extended for precise instance segmentation of individual trees. Furthermore, we demonstrated that the quality of the single-tree delineation could be significantly optimized. All experiments were compared using a state-of-the-art instance segmentation Mask R–CNN and four baseline methods for single-tree segmentation. In detail, using reference data collected by field measurements and visual interpretation, the experiments showed that the two instance segmentations hardly differ in accuracy and perform significantly better than baseline methods. The best results were achieved using the *CIR* channel combination. Interestingly, the inclusion of height information did not increase the accuracy. The superiority of instance segmentation was particularly evident in the quality of the segmentation that was up to 16% better than baseline methods. Moreover, we could not show that images from lidar data increased the accuracy despite tree stems being visible in the lidar data in some parts of the study area. Note that in the case of a high point density, lidar images could be used along with a separate second backbone to generate additional features that represent the stem information and are combined with the feature set of the first backbone. Furthermore, a lidar flight mission conducted in a leaf-off condition increases the point density below the tree canopy. Modifications of the DETR methodology to decrease the number of parameters could be achieved by reducing the number of the encoder–decoder layers. In addition, the loss calculation could be improved using the complete IoU (Zheng et al., 2021). Deformable DETR (Zhu et al., 2020) could improve the performance in small trees by using a multi-scale deformable attention module. Finally, sufficiently trained networks using our approach could be applied to generate reference data for large-scale forest inventories in the same forest area using aerial or satellite images with lower resolution than UAV-imagery.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the Research Department Life Science Systems of the Technical University Munich for the support sensing data and cooperation, especially Dr. rer. nat. Thomas Schneider for supporting us with sensing data. This research was partially funded by the Federal Ministry of Education and Research (BMBF) under the contract number 13FH004IX6 and by the Deutsche Bundesstiftung Umwelt under der contract number 37681/01.

APPENDIX

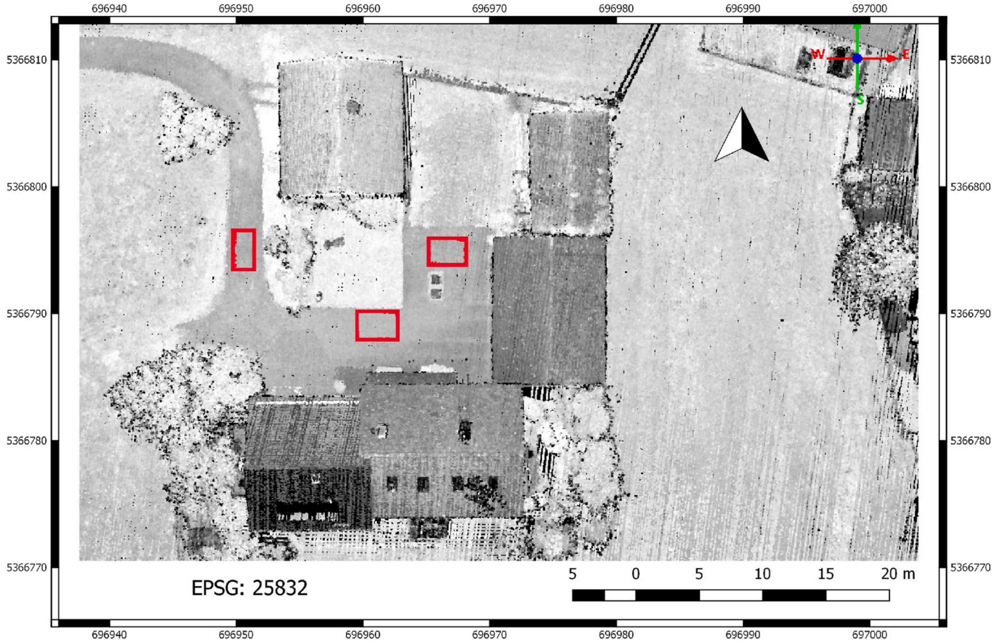
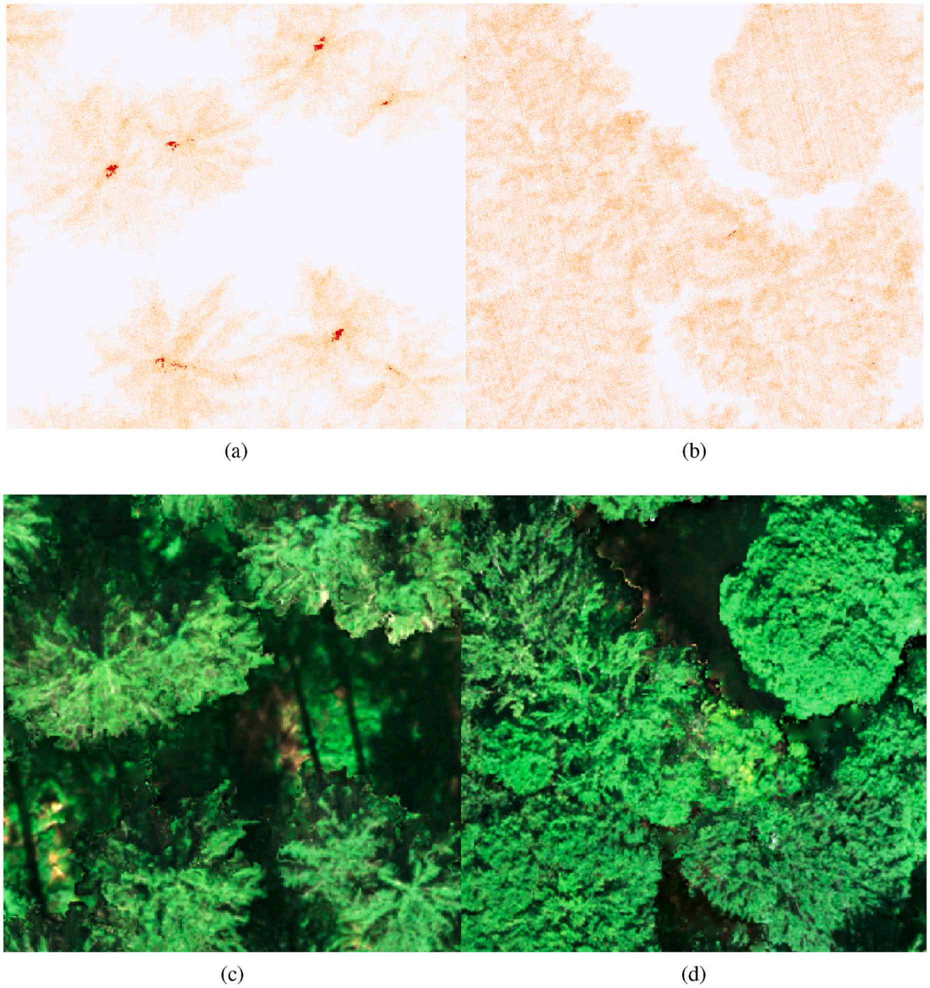


Fig. A1. Three small concrete areas (red polygons) providing mean values I_n for calibration of lidar intensity.

Table A2
Control parameters of baseline methods Li et al. (2012) (=Study #1), Silva et al. (2016) (=Study #2), Reitberger et al. (2009) (=Study #3) and Roussel and Auty (2022) (=Study #4) for tree segmentation applied to lidar data. Note that all other control parameters were set to default.

Study #1			Study #2		
R1[m]	dt1[m]	dt2[m]	chm[m]	window size [m]	max_cr[m]
2.0	1.5	2.5	0.5	2.0	0.6
	Study #3			**Study #4**	
NCut	σ_{xy}	σ_z	th_tree	tol	ext
0.09	1.35	11.0	2.0	0.5	2.0


(a) (b)

(c) (d)

Fig. A3. Sections of mainly coniferous (a,c) and deciduous (b,d) trees. Lidar-based layer *P_DENSE* containing tree stem locations visible as red dots (a) and without clearly visible tree stems (b). Corresponding orthophotos are shown in (c) and (d).

References

BayesMap, 2018. Stripalign. http://bayesmap.com/products/bayesstripalign/, 2021-12-08.

G Braga, J.R., Peripato, V., Dalagnol, R., P Ferreira, M., Tarabalka, Y., O C Aragão, L.E., F. de Campos Velho, H., Shiguemori, E.H., Wagner, F.H., 2020. Tree crown delineation algorithm based on a convolutional neural network. Rem. Sens. 12 (8) https://doi.org/10.3390/rs12081288. ISSN 2072-4292.

Briechle, S., Krzystek, P., Vosselman, G., 2021. Silvi-net – a dual-cnn approach for combined classification of tree species and standing dead trees from remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 98, 102292 https://doi.org/10.1016/j.jag.2020.102292. ISSN 1569-8432.

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, ISBN 978-3-030-58452-8, pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

Chadwick, A.J., Goodbody, T.R.H., Coops, N.C., Hervieux, A., Bater, C.W., Martens, L.A., White, B., Röeser, D., 2020. Automatic delineation and height measurement of regenerating conifer crowns under leaf-off conditions using uav imagery. Rem. Sens. 12 (24) https://doi.org/10.3390/rs12244104.

Chen, G., Shang, Y., 2022. Transformer for tree counting in aerial images. Rem. Sens. 14 (3) https://doi.org/10.3390/rs14030476. ISSN 2072-4292.

Dalponte, M., Coomes, D.A., 2016. Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. Methods Ecol. Evol. 7 (10), 1236–1245. https://doi.org/10.1111/2041-210X.12575.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. Imagenet, 2009. A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Dersch, S., Heurich, M., Krueger, N., Krzystek, P., 2021. Combining graph-cut clustering with object-based stem detection for tree segmentation in highly dense airborne lidar point clouds, 207–222 ISPRS J. Photogrammetry Remote Sens. 172. https://doi.org/10.1016/j.isprsjprs.2020.11.016. ISSN 0924-2716.

Dersch, S., Schöttl, A., Krzystek, P., Heurich, M., 2022. Novel single tree detection by transformers using uav-based multispectral imagery. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 43, 981–988. https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-981-2022.

Diez, Y., Kentsch, S., Fukuda, M., Caceres, M.L.L., Moritake, K., Cabezas, M., 2021. Deep learning in forestry using uav-acquired rgb data: a practical review. Rem. Sens. 13 (14) https://doi.org/10.3390/rs13142837.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

Hao, Z., Lin, L., Post, C.J., Mikhailova, E.A., Li, M., Chen, Y., Yu, K., Liu, J., 2021. Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (mask r-cnn). ISPRS J. Photogrammetry Remote Sens. 178, 112–123. https://doi.org/10.1016/j.isprsjprs.2021.06.003. ISSN 0924-2716.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. *CoRR*, abs 1512, 03385. https://doi.org/10.1109/CVPR.2016.90. URL.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, 2980–2988. In: 2017 IEEE International Conference on Computer Vision. ICCV. https://doi.org/10.1109/ICCV.2017.322.

Heurich, M., 2006. Evaluierung und Entwicklung von Methoden zur automatisierten Erfassung von Waldstrukturen aus Daten flugzeuggetragener Fernerkundungssensoren. PhD thesis. Technische Universität München.

Heurich, M., 2008. Automatic recognition and measurement of single trees based on data from airborne laser scanning over the richly structured natural forests of the bavarian forest national park. For. Ecol. Manag. 255 (7), 2416–2433. https://doi.org/10.1016/j.foreco.2008.01.022 (Large-scale experimentation and oak regeneration).

Höfle, B., Pfeifer, N., 2007. Correction of laser scanning intensity data: data and model-driven approaches. ISPRS J. Photogrammetry Remote Sens. 62 (6), 415–433. https://doi.org/10.1016/j.isprsjprs.2007.05.008. ISSN 0924-2716.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on convolutional neural networks (cnn) in vegetation remote sensing, 24–49 ISPRS J. Photogrammetry Remote Sens. 173. https://doi.org/10.1016/j.isprsjprs.2020.12.010. ISSN 0924-2716.

Krzystek, P., Serebryanyk, A., Schnörr, C., Červenka, J., Heurich, M., 2020. Large-scale mapping of tree species and dead trees in Šumava national park and bavarian forest national park using lidar and multispectral imagery. Rem. Sens. 12 (4) https://doi.org/10.3390/rs12040661. ISSN 2072-4292.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Nav. Res. Logist. Q. 2 (1–2) https://doi.org/10.1002/nav.3800020109, 83–97.

LasTools, 2021. Geocode. http://www.geolas.com/Downloads/Geocode_Overview_EN_Rev_20190908.pdf, 2021-12-08.

Latifi, H., Fassnacht, F.E., Müller, J., Tharani, A., Dech, S., Heurich, M., 2015. Forest inventories by lidar data: a comparison of single tree segmentation and metric-based methods for inventories of a heterogeneous temperate forest. Int. J. Appl. Earth Obs. Geoinf. 42, 162–174. https://doi.org/10.1016/j.jag.2015.06.008. ISSN 1569-8432.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539. ISSN 1476-4687.

Li, W., Guo, Q., Jakubowski, M., Kelly, M., 2012. A new method for segmenting individual trees from the lidar point cloud. Photogramm. Eng. Rem. Sens. 78 (75–84) https://doi.org/10.14358/PERS.78.1.75.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.

Lindner, M., Maroschek, M., Netherer, S., Kremer, A., Barbati, A., Garcia-Gonzalo, J., Seidl, R., Delzon, S., Corona, P., Kolström, M., Lexer, M.J., Marchetti, M., 2010. Climate change impacts, adaptive capacity, and vulnerability of european forest ecosystems. For. Ecol. Manag. 259 (4), 698–709. https://doi.org/10.1016/j.foreco.2009.09.023. Adaptation of Forests and Forest Management to Changing Climate.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C. Ssd, 2016. In: Single Shot Multibox Detector, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

MetaShape. Professional edition, 2010. https://www.agisoft.com/features/professional-edition/, 2022-03-29.

RIEGL. minivux-1uav, 2020. http://www.riegl.com/products/unmanned-scanning/riegl-minivux-1uav/, 2021-12-08.

Novatel. Intertial explorer, 2018. https://de.calameo.com/read/00191579637d5b7ede95d?authid=4AxY37U3TTL8, 2021-12-08.

PrimaVision, 2022. 3d tree segmentation from point clouds (lidar, dsm) for forest inventory. http://primavision-tec.de/products/prod_tree-finder, 2022-08-19.

RedEdge-MX, 2020. Red edge mx. https://micasense.com/rededge-mx/, 2022-03-29.

Redmon, J., Farhadi, A., 2018. Yolov3: an Incremental Improvement. *CoRR*, abs/1804.02767. https://doi.org/10.48550/arXiv.1804.02767. URL.

Redmon, J., Farhadi, A., 2016. YOLO9000: Better, Faster, Stronger. *CoRR*, abs/1612.08242. https://doi.org/10.48550/arXiv.1612.08242. URL.

Reid, W.V., et al., 2005. Millennium Ecosystem Assessment.

Reitberger, J., Schnörr, C., Krzystek, P., Stilla, U., 2009. 3d segmentation of single trees exploiting full waveform lidar data. ISPRS J. Photogrammetry Remote Sens. 64 (6), 561–574. https://doi.org/10.1016/j.isprsjprs.2009.04.002. ISSN 0924-2716.

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 (6), 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. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658–666. https://doi.org/10.1109/CVPR.2019.00075.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Roussel, J.-R., Auty, D., 2022. https://rdrr.io/github/Jean-Romain/lidR/man/its_watershed.html. R package version 4.0.1.

Roussel, J.-R., Auty, D., Coops, N.C., Tompalski, P., Goodbody, T.R., Meador, A.S., Bourdon, J.-F., de Boissieu, F., Achim, A. lidr, 2020. An r package for analysis of airborne laser scanning (als) data. Rem. Sens. Environ. 251, 112061 https://doi.org/10.1016/j.rse.2020.112061. ISSN 0034-4257.

Seidl, R., Schelhaas, M.-J., Rammer, W., Verkerk, P.J., 2014. Increasing forest disturbances in europe and their impact on carbon storage. Nat. Clim. Change 4 (9), 806–810. https://doi.org/10.1038/nclimate2318.

Sentera6X. Sentera6x, 2019. https://sentera.com/wp-content/uploads/2022/03/6X-Multispectral-Sensor-Flyer.pdf, 2022-08-05.

Silva, C.A., Hudak, A.T., Vierling, L.A., Loudermilk, E.L., O'Brien, J.J., Hiers, J.K., Jack, S.B., Gonzalez-Benecke, C., Lee, H., Falkowski, M.J., Khosravipour, A., 2016. Imputation of individual longleaf pine (pinus palustris mill.) tree attributes from field and lidar data. Can. J. Rem. Sens. 42 (5), 554–573. https://doi.org/10.1080/07038992.2016.1196582.

Strîmbu, V.F., Strîmbu, B.M., 2015. A graph-based segmentation algorithm for tree crown extraction using airborne lidar data. ISPRS J. Photogrammetry Remote Sens. 104 (30–43) https://doi.org/10.1016/j.isprsjprs.2015.01.018. ISSN 0924-2716.

TerraSolid, 2020. Terra scan. https://terrasolid.com/products/terrascan/, 2021-12-08.

Thom, D., Seidl, R., 2016. Natural disturbance impacts on ecosystem services and biodiversity in temperate and boreal forests. Biol. Rev. 91 (3), 760–781. https://doi.org/10.1111/brv.12193.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, ume 30. Curran Associates, Inc., pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

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, 10 Forestry: Int. J. Financ. Res. 85 (1), 27–40. https://doi.org/10.1093/forestry/cpr051.

Weinstein, B.G., Marconi, S., Bohlman, S., Zare, A., White, E., 2019. Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. Rem. Sens. 11 (11) https://doi.org/10.3390/rs11111309. ISSN 2072-4292.

White, J.C., Coops, N.C., Wulder, M.A., Vastaranta, M., Hilker, T., Tompalski, P., 2016. Remote sensing technologies for enhancing forest inventories: a review. Can. J. Rem. Sens. 42 (5), 619–641. https://doi.org/10.1080/07038992.2016.1207484.

Windrim, L., Bryson, M., 2020. Detection, segmentation, and model fitting of individual tree stems from airborne laser scanning of forests using deep learning. Rem. Sens. 12 (9) https://doi.org/10.3390/rs12091469. ISSN 2072-4292.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Yao, W., Krzystek, P., Heurich, M., 2012. Tree species classification and estimation of stem volume and dbh based on single tree extraction by exploiting airborne full-waveform lidar data. Rem. Sens. Environ. 123, 368–380. https://doi.org/10.1016/j.rse.2012.03.027.

YellowScanVoyager, 2022. Voyager. https://www.yellowscan-lidar.com/products/voyager/, 2022-08-05.

Yu, X., Hyyppä, J., Holopainen, M., Vastaranta, M., 2010. Comparison of area-based and individual tree-based methods for predicting plot-level forest attributes. Rem. Sens. 2 (6), 1481–1495. https://doi.org/10.3390/rs2061481.

Zhang, X., 2022. Simple understanding of mask rcnn. https://alittlepain833.medium.com/simple-understanding-of-mask-rcnn-134b5b330e95, 2022-08-05.

Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., Zuo, W., 2021. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans. Cybern. 52 (8), 8574–8586. https://doi.org/10.48550/ARXIV.2005.03572.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable DETR: Deformable Transformers for End-To-End Object Detection. *CoRR*, abs/2010.04159. https://doi.org/10.48550/arXiv.2010.04159.