NOVEL SINGLE TREE DETECTION BY TRANSFORMERS USING UAV-BASED MULTISPECTRAL IMAGERY

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

[a] Dept. of Geoinformatics, Munich University of Applied Sciences, 80333 Munich, Germany
(sebastian.dersch, peter.krzystek)@hm.edu
[b] Faculty of Environment and Natural Resources, University of Freiburg, Germany
marco.heurich@wildlife.uni-freiburg.de
[c] Bavarian Forest National Park, Dept. of Visitor Management and National Park Monitoring, 94481 Grafenau, Germany
marco.heurich@npv-bw.bayern.de
[d] Dept. of Electrical Engineering and Information Technology, Munich University of Applied Sciences, 80333 Munich, Germany
alfred.schoettl@hm.edu
[e] Institute for Applications of Machine Learning and Intelligent Systems, Munich University of Applied Sciences
80333 Munich, Germany

Commission II, WG II/10

KEY WORDS: Neural networks, transformer, single tree detection, multispectral imagery, UAV.

ABSTRACT:

Single tree detection has been a major research topic concerning automatic forest inventory using remote sensing data. Recently, deep learning-based approaches in remote sensing forestry have gained attention because of the prospect of improved accuracy. In this study, we present a novel tree detection method based on the detection transformer (DETR), which applies a transformer in combination with a pre-trained convolutional neural network to detect individual trees using high-resolution multispectral imagery. The test site (Kranzberg Forest Roof Experiment - KROOF) is located in Bavaria, north of Munich, and is characterised by a mixed forest which consists of large groups of European beeches (Fagus sylvatica) surrounded by Norway spruces (Picea abies). The image data were acquired with a MicaSense RedEdge-MX Dual camera mounted to UAV. Two flight mission were conducted at an altitude of around 85 m with a flight speed of 5 m/sec, resulting in a ground resolution of about 5 cm. 125 trees were surveyed by tacheometric means in the field for testing, and 1390 trees were labelled by visual interpretation of the multispectral imagery for training and validation. The novel tree detection method based on DETR shows promising results and outperforms the standard, well-known object detection method YOLOv4 in mixed and deciduous test plots. More detailed, F1-scores were evaluated for coniferous plot at 83%, for mixed plot at 86% and for deciduous plot at 71%. The corresponding figures for YOLOv4 are 87% coniferous, 65% mixed and 67% deciduous. In terms of accuracy, DETR is inferior by 6% in coniferous plot, however superior by 28% and 5% in mixed and deciduous plot, respectively. Compared to YOLOv4, we found that DETR sometimes failed to detect small coniferous trees. Moreover, both deep learning-based methods tend to over-detect single trees in deciduous test areas. In sum, transformer-based tree detection shows great potential to improve single tree detection.

1. INTRODUCTION

Forests are an essential part of our environment, providing critical ecosystem services, such as carbon storage, nutrient cycling, drinking water supply and air purification. Moreover, they offer recreational opportunities and host a large proportion of Earth´s biodiversity. Forest loss, global change and an unsustainable management are threatening forest ecosystems in an unpreceded manner. A better knowledge of the condition of the forests is a prerequisite for sound management, for which forest inventories form an important basis. Here, remote sensing methods come into play as they can acquire this information over large areas at a much lower cost in comparison to conventional methods (Krzystek et al., 2020).
Forest inventories, as part of sustainable forest management, are usually conducted on small sample plots (less than 1% of the area) with intensive terrestrial measurements, which survey individual tree attributes and derive statistical indicators for the surveyed areas. An areal wide collection of forest structure parameters down to single tree information can only be done by using remote sensing methods and offers an added value for the areal monitoring of forest structures (Latifi et al., 2015). Using high-resolution remote sensing data and innovative AI methods, this information can be collected over large areas at a much lower cost (Latifi and Heurich, 2019).

2. RELATED WORK

In recent years, the use of deep neural networks (DNN), such as segmentation and classification algorithms, has attracted a great deal of interest as they outperform standard machine learning approaches in various tasks (Voulodimos et al., 2018). The main advantage of many DNNs is representation learning, which characterises automatic feature extraction as part of the training process (LeCun et al., 2004). However, single tree detection and segmentation via deep learning are more challenging and only a few approaches apply instance segmentation that imbed two-stage object detectors to delineate single trees using lidar data (Windrim and Bryson, 2020) or multispectral imagery (G. Braga et al., 2020). In another study, a tree detection method based on the single-stage detector RetinaNet (Lin et al., 2017)

[*] Corresponding author.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-981-2022 | © Author(s) 2022. CC BY 4.0 License.

using RGB imagery is presented (Weinstein et al., 2019). The model is initially trained by tree segments provided by a lidar-based segmentation and is fine-tuned using manually labeled segments. When applied in an open forest area, the approach outperforms two baseline methods (Silva et al., 2016), (Li et al., 2012).

The novel use of transformers is promising (Parmar et al., 2018), which are deep learning building blocks using the mechanism of self-attention (Vaswani et al., 2017). In this work, we aim to detect single trees in high-resolution RGB true orthophotos (TDOPs) using a novel transformer approach detection transformer (DETR) (Carion et al., 2020). In an interesting study, a similar procedure was applied in the field of bioinformatics (Prangemeier et al., 2020). It was successfully shown that cells in microstructures can be detected and classified using microscope imagery with the help of transformers. In remote sensing, change detection in residential areas was conducted, reporting accuracy improvements compared to baseline architectures such as U-Net (Chen et al., 2021). Recently, a study presented a new deep learning model called density transformer (DENT) for automatic tree counting from aerial images (Chen and Shang, 2022). The architecture is similar to DETR in (1) using a convolutional neural network for extraction of visual features and (2) providing contextual image information with the help of conventional transformer encoder in a multi-head attention mechanism. The encoder gives input for two separate feed-forward networks: one that generates a tree density map and another that counts trees. DENT outperforms most of the other deep learning-based methods such as Faster R-CNN (Ren et al., 2015) and YOLOv3 (Redmon and Farhadi, 2018).

To the authors' best knowledge, so far no experiments have been carried out using this new far-reaching deep learning-based object detection method to detect single trees in a high-resolution TDOP in the context of forest inventory. In order to demonstrate the potential of the transformer-based method, the results were compared with a well-known one-stage object detection method called You Only Look Once v4 (YOLOv4) (Bochkovskiy et al., 2020).

3. MATERIAL

3.1 Study area

Our experiments were conducted close to the Kranzberg Forest Roof Experiment (KROOF) research site, located at 11°39‘42” E, 48°25‘12” N, approximately 35 km northeast of Munich. The forest around the KROOF research site is under administration of the Bayerische Staatsforsten. Most of the mixed forest is characterised by large groups of beeches surrounded by spruces. Tree heights vary between 19 m and 36 m with a stem density of around 200–300 trees/ha.

For the evaluation, field measurements were conducted to generate reference data. For trees with a breast height diameter (BHD) greater than 15 cm, the tree positions were measured by tacheometric means with an accuracy of less than 2 cm. The BHD was conventionally determined using a caliper. The first plot (Figure 1, Plot #1) is characterised by dominant coniferous trees and some understory trees as well. The second plot (Figure 1, Plot #2) is more diverse, composed of 60% coniferous and 40% deciduous trees. The third plot (Figure 1, plot #3) is dominated by deciduous trees which make up 76% of the area. The variety also refers to the size and the age of the occurring trees. Table 1 shows the plot characteristics. Since a 2D data based method is used, only dominant trees and trees recognisable in the TDOP were used for the accuracy assessment. Figure 1 shows test plots #1 and #2 superimposed on the TDOP of the August 2020 flight. Test plot #3 is shown on the data set flown in July 2021.

3.2 Data acquisition and preparation

3.2.1 Aerial multispectral data In August 2020 and July 2021, multispectral images were collected using a RedEdge MX Dual camera (MicaSense, 2022) attached to a remotely piloted hexacopter (DJI M 600 Pro). The camera system captures ten channels (spectral range 475 – 842 nm) with a horizontal field of view of (HFOV) of 47.2°, which corresponds to a focal length of 5.5 mm. A downwelling light sensor provided accurate ambient light calibration. Images of a calibration panel were taken for radiometric calibration. The flight speed was 5 m/sec above ground. The end lap and side lap of the image block were 90% and 60%, respectively. For the two missions, the flight heights were 90 m and 80 m, resulting in ground sample distances (GSDs) of 5.93 cm and 5.3 cm. For postprocessing of the imagery, structure-from-motion (SFM) software was used to generate TDOPs (MetaShape, 2022). The processing steps consisted of (1) radiometric calibration of imagery (2), bundle adjustment, (3) point cloud generation and (4) generation of an orthomosaic. The exported TDOPs had a cell size of 5 cm containing ten channels captured by the camera. Table 2 provides an overview of the photogrammetric campaign.

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	deciduous
Tree heights (m)	19-34	20-34	19-34
Images	19	22	15

Table 1. Parameters of reference plots.

Multispectral camera	RedEdge MX
SFM - Software	MetaShape
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

Table 2. Flight parameters of aerial image acquisition and software packages used.

Parameter	Training	Validation
Size (m^2)	48175	8500
Trees	1167	223
Trees/ha	243	262
Forest type	mixed	mixed
Tree heights (m)	19-34	20-34
Images	433	72

Table 3. Parameters of training and validation areas captured by visual inspection.

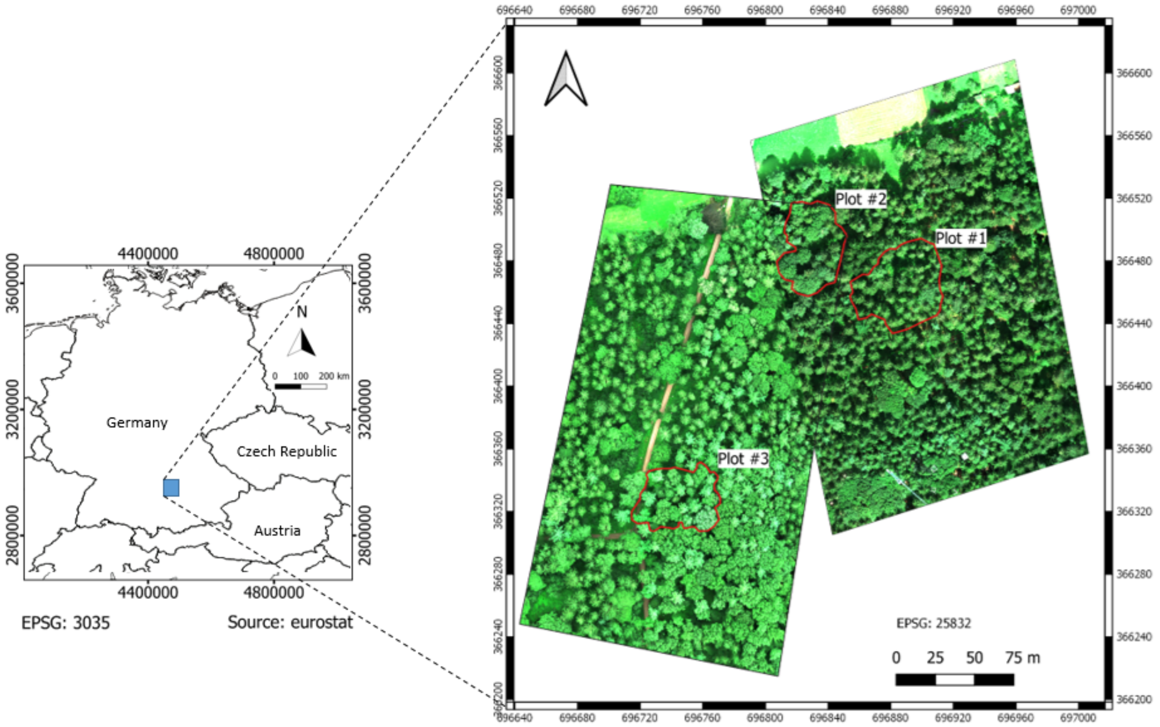


Figure 1. Research site KROOF as RGB TDOP showing three test plots #1, #2, and #3. The remaining area was used for training and validation.

3.2.2 Field survey The goal of the field campaign was to measure tree positions as precisely as possible in order to generate accurate test data (see 3.2.3). Due to the expected shading effects in dense forest areas using global navigation satellite system (GNSS) systems, a survey campaign was conducted in April 2021. First, a traverse was measured in the area of plots #1, #2 and #3. The traverse included seven polygon stations and was georeferenced using three geodetic points. The Trimble R12i GNSS system and the Leica TCRP1203+ total station were used as instruments. Afterwards, tree positions were surveyed from the polygon points by tacheometric means. The BHD of each tree greater than 15 cm was also measured using a caliper, and the tree group was also documented. In summary, 55 trees, 36 trees and 34 trees were surveyed in plots #1, #2 and #3, respectively (see also Table 1). The estimated accuracy of the tree positions was less than 10 cm.

3.2.3 Labeling of tree crowns The training and test reference data are provided in the form of enclosing bounding boxes. For this purpose, the TDOP is used for visualisation. For the labeling of the training data, tree segments were defined in the TDOP. The tree segments of the test data are also determined using the TDOP and additionally linked with tree positions derived from the field measurements. Figure 2 shows an example of labeled trees with corresponding bounding boxes and tree positions.

4. METHODS

4.1 DETR

The deep learning-based method DETR considers object detection as a direct set prediction problem. This approach processes global image information by the transformer mechanism (Vaswani et al., 2017) and eliminates the need for several subtasks that require prior knowledge about the problem, such as anchor generation. Predictions are determined directly and in parallel using a small set of learned object queries. The relationship between objects and the global image context is a key factor in this process.

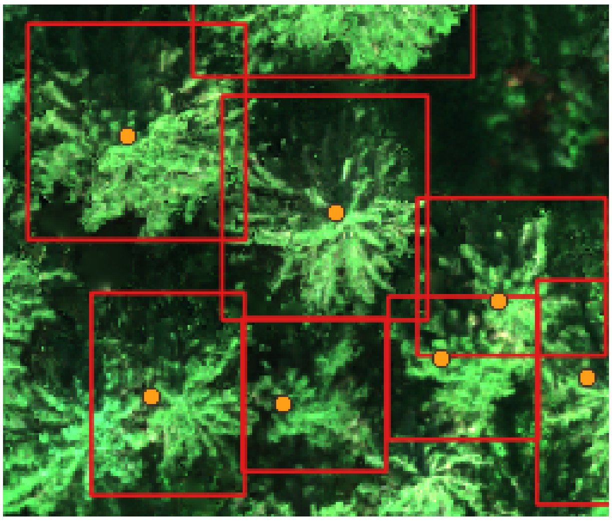


Figure 2. Example of a TDOP with labeled bounding boxes colored in red and the corresponding tree positions as points colored in orange.

The overall DETR architecture with the key elements is illustrated in Figure 3. First, the features are extracted using ResNet-50 (He et al., 2015). Afterwards, the positional encodings are added up element by element to the CNN features. Finally,

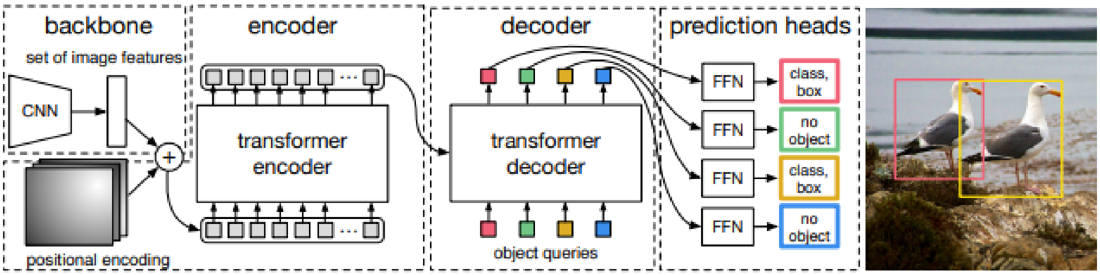


Figure 3. Illustration of design and functionality of the object detection method DETR. Image source of (Carion et al., 2020).

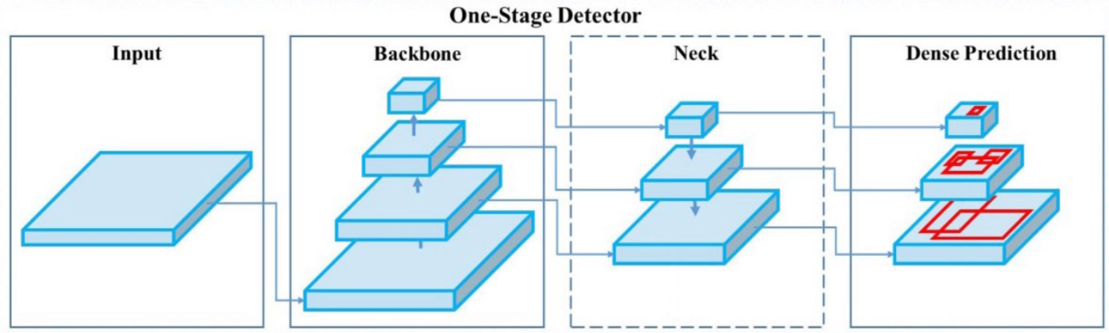


Figure 4. Illustration of design and functionality of the object detection method YOLOv4. Image source of (Bochkovskiy et al., 2020).

the result is transferred to a transformer encoder followed by a transformer decoder, which generates N object queries. The last step classifies bounding boxes and classes using an feed-forward network (Carion et al., 2020). Beside the architecture, finding and evaluating ground truth and predicted boxes plays a crucial role. The definition of a set prediction loss, which includes a unique matching procedure between ground truth boxes and a larger set of predicted boxes, has been determined efficiently by bipartite matching using the Hungarian algorithm (Kuhn, 1955). The total loss is a combination of the matching loss and the Hungarian loss, which includes a linear combination of the generalised intersection over union (IoU) loss (Rezatofighi et al., 2019) and the L1 loss.

This architecture offers several advantages. No previous information about anchors is needed and global information can be processed due to the transformer mechanism. However, on the other hand this workflow has issues detecting small objects compared to the faster R-CNN (Ren et al., 2015) and converges slower than comparable object detection methods. The new design based on transformers and bipartite matching in the area of object detection and the good extensibility of the workflow offers the possibility for adaptations in different fields. For example, the authors of Deformable DETR (Zhu et al., 2020) extended the existing workflow so that Deformable DETR detects smaller objects better and requires a factor of 10 less training epochs. The authors of Dynamic DETR (Dai et al., 2021) have significantly reduced the number of training epochs and achieved improved performance by introducing a dynamic encoder that reduces the quadratic computational complexity of the self-attention module in transformer encoders.

4.2 YOLOv4

YOLOv4 is the fourth evolutionary step of the original You Only Look Once (YOLO) (Redmon et al., 2015) released in a flexible research framework called darknet. The original version included the first neural net approach that could generate all bounding boxes and class labels parallel in one inference step using an end-to-end network. Historically, YOLO has undergone several improvement iterations with YOLOv2 (Redmon and Farhadi, 2016), YOLOv3 (Redmon and Farhadi, 2018) and YOLOv4 (Bochkovskiy et al., 2020). YOLOv4 achieves state-of-the-art detection accuracy in roughly realtime. The architecture illustrated in Figure 4 shows the essential components of the workflow.

First, features are extracted from images using the feature extractor CSPdarknet53. Here, cross-stage partial connections are attached to darknet53 from YOLOv3. As feature aggregator, spatial pyramid pooling is utilised as it increases the receptive field and differs the most important features. Then, instead of the feature pyramid network in YOLOv3, the path aggregation network is utilized. The original YOLOv3 network was used as head to generate bounding boxes and class labels.

Beside the architecture, two strategies have been introduced. One of these strategies is called bag of freebies, which does not require any additional computing power and uses data augmentation, such as mosaic or cutmix. The other strategy, bag of specials, contains improvement modules for inference (e.g. mish activation) (Misra, 2019).

5. EXPERIMENTS

5.1 Experimental setup

Due to the architecture of DETR and YOLOv4, we preprocessed the training, validation and test image data in 50% overlapping tiles of 512 x 512 pixels. For the training process, the training data was split into 80% training and 20% validation. Table 1 and 3 show the number of images used for training, validation and testing. Three models each were trained for DETR and YOLOv4 with varying random number generator seeds to check the reproducibility of the results. In the case of DETR, no remarkable variation of the results were found. Instead, YOLOv4 exhibited a wider deviation of the results. We therefore computed mean values of statistical parameters accuracy, precision, recall and F1-score (See Section 5.4) in the respective plots #1, #2, and #3. As a result of the 50% overlap in the images, overlapping bounding boxes were predicted during testing. At the image edges, small tree fragments sometimes occurred. Therefore, post-processing was necessary to filter the small bounding box fragments using a threshold value (e.g. 20 m^2). Subsequently, a non-maximum suppression was applied using an IoU threshold of 0.3. A workstation equipped with 256 GB RAM, a Nvidia RTX 8000 GPU, and an AMD Ryzen Threadripper 3970X processor was used.

5.2 Configuration of DETR

The DETR Python implementation of Hugging Face (Wolf et al., 2020) was used and the default configuration of DETR was applied with an adjusted learning rate of 1e-7, a backbone learning rate of 1e-6, a weight decay of 1e-8 and a batch size of 12. A pre-trained model based on the common objects in context (COCO) detection dataset (Lin et al., 2014) was used because of the limited amount of training data. This required a constant parameter value object queries (optimized for the COCO dataset) to be fixed to 100. Within a period of 500 epochs, we applied early stopping to train and validate the model, thereby mitigating overfitting effects. To achieve this, the model with the highest validation mean average precision @IoU=0.50 (mAP) was selected first. It was checked whether the validation loss was within the range of a minimum. Due to the 50% overlapping test images and the object queries parameter, the bounding box fragments with high confidence scores at the edges of the test images were eliminated in an intermediate step.

5.3 Configuration of YOLOv4

In this work, the YOLOv4 implementation for Windows computers was used (Bochkovskiy et al., 2020). The configuration was adapted for this data set by setting the number of training steps to 7200 and using a batch size of 64. Data augmentation (crop, rotation, flip, hue, saturation, exposure, aspect, cutmix, mixup, mosaic and blur) was also applied. A pre-trained model based on the COCO detection data set (Lin et al., 2014) was selected for transfer learning. The model with the maximum mean average precision (Everingham et al., 2010) value was selected within 7200 training steps.

5.4 Accuracy assessment

In order to determine the quality of the tree detection, the following metrics were used. First, accuracy, precision, recall and F1-score were taken to identify the performance of the results. Detected trees that could be assigned to a reference bounding box with at least an IoU of 50 % were taken as successful detected trees (true positives). If no assignment to a reference bounding box was found for detected trees, they were categorised as false positives. Furthermore, reference trees, which could not be matched to any detected tree, were marked as false negatives. The IoU describes the quality of the overlap and is defined as the ratio of the common area and the combined area of the bounding boxes A and B. Equations 1, 2, 3, 4 and 5 show how the described parameters are calculated, whereby TP, FP, FN and F1 are denoted as true positives, false positives, false negatives and F1-score.

$$accuracy = \frac{TP}{TP + FP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

6. RESULTS AND DISCUSSION

Tree detection results using DETR and YOLOv4 methods are summarised in Table 4. DETR clearly outperforms YOLOv4 in mixed plot #2 and deciduous plot #3. Interestingly, mixed plot #3 shows a significant difference of more than 20% in terms of F1-score. The F1-score in deciduous plot #3 is lower with 4%. In contrast to plot #2 and #3, DETR deteriorates by 4% F1-score in coniferous plot #1. Note that the accuracy values in Table 4 fully confirm the trend of the F1-score.

Across all three test plots, both methods have problems with over-segmentation. The effect is clearly distinctive in mixed plot #2. This is reflected in a 23% lower recall for YOLOv4. Figure 6 shows a deciduous crown with a diameter of 15 m completely detected by DETR and split-up into two boxes by YOLOv4. Our explanation for this is that the training data contain mainly medium-sized trees. Therefore, to reduce the over-segmentation effect, more larger tree crowns should be included in the training.

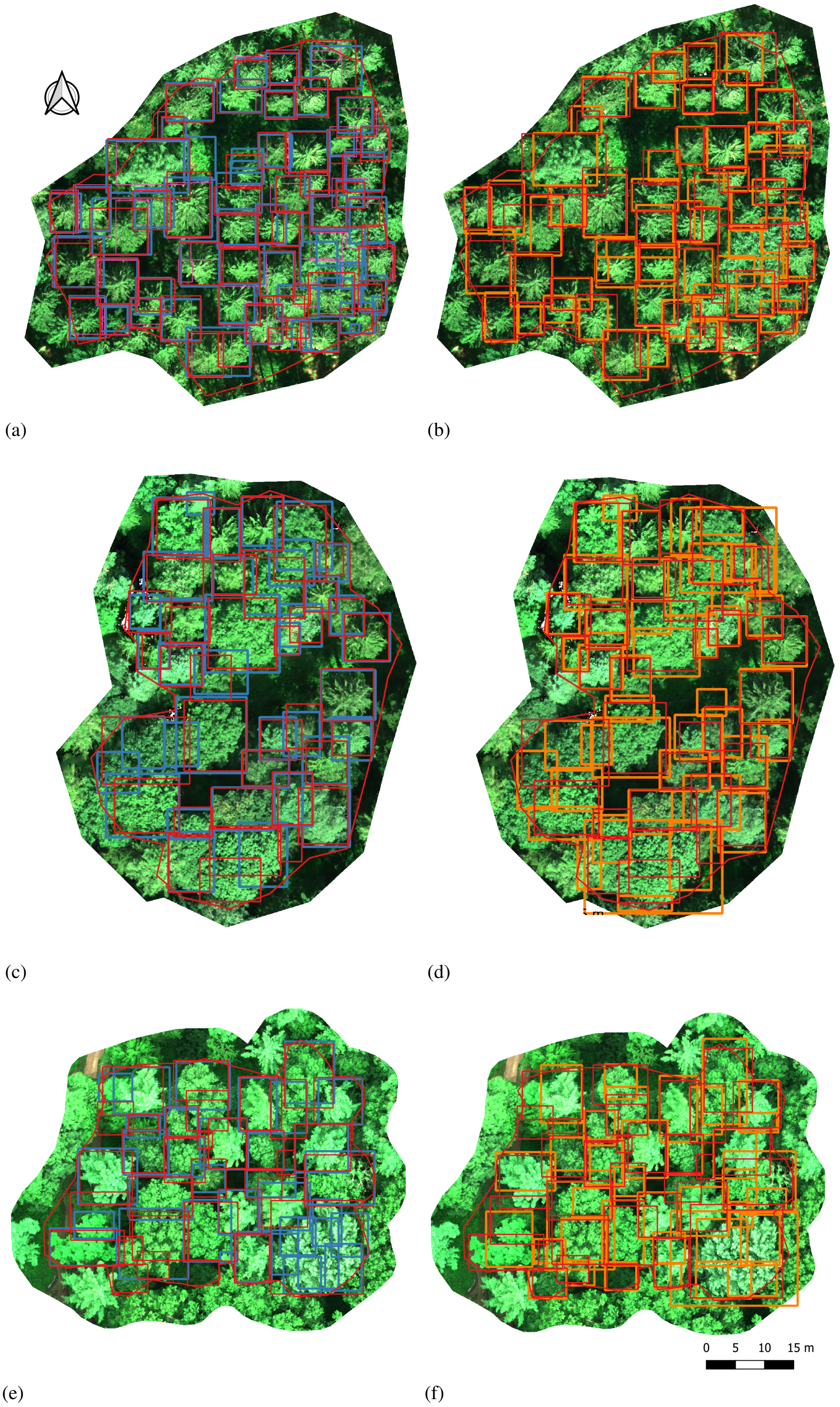
Furthermore, we notice that DETR obviously detects smaller trees worse than YOLOv4. This is especially noticeable in plot #1, where a total of seven small trees with a crown diameter of less than 5 m are located. More detailed, DETR detects only one tree, however YOLOv4 is able to successfully find four trees in this plot. To clarify this, Figure 7 shows a sample subarea of plot #1. In Figure 7a, we notice that DETR detects two small trees. However, these are false positives because of a too low IoU value. Instead, YOLOv4 successfully detects two trees in this subarea (See Figure 7b).

In conclusion, DETR has apparently problems to detect smaller trees, which is also reflected in poorer results in plot #1. This can be explained mainly by the fact that DETR generally detects smaller objects worse than object detectors such as YOLOv4 or Faster-RCNN, which normally use higher resolution feature maps. This disadvantage of DETR was recently compensated by an extension called Deformable DETR, which achieves significantly better detection results for small objects (Zhu et al., 2020).

Finally, we compare our results with the study (Weinstein et al., 2019) that utilizes the NEON woody vegetation dataset (National Ecological Observatory Network (NEON), 2022) with images (GSD = 0.1 m) acquired at the San Joaquin Experimental Range in California. The forest area is characterised as open forest comprising the predominant tree species live oak (Quercus agrifolia), blue oak (Quercus douglasii) and foothill

	coniferous plot #1				mixed plot #2				deciduous plot #3			
	Acc.	F1-score	Prec.	Rec.	Acc.	F1-score	Prec.	Rec.	Acc.	F1-score	Prec.	Rec.
DETR	71	83	81	85	76	86	84	89	55	71	71	71
YOLOv4	77	87	84	89	48	65	57	75	50	67	63	71

Table 4. Results of tree detection with DETR and YOLOv4 in test plots #1, #2 and #3. Numbers in percent.



(a) (b)

(c) (d)

(e) (f)

Figure 5. Results for object detection with DETR and YOLOv4. Reference data plot area surroundings are in red. Predicted bounding boxes are in blue (DETR) and orange (YOLOv4) respectively. a) Result of DETR for coniferous plot #1. b) Result of YOLOv4 for coniferous plot #1. c) Result of DETR for mixed plot #2. d) Result of YOLOv4 for mixed plot #2. e) Result of DETR for deciduous plot #3. f) Result of YOLOv4 for deciduous plot #3.

pine (Pinus sabiniana). The study reports a detection accuracy of 69% recall and 61% precision. Comparison with our study is difficult due to differences in the characteristics of the forest area.

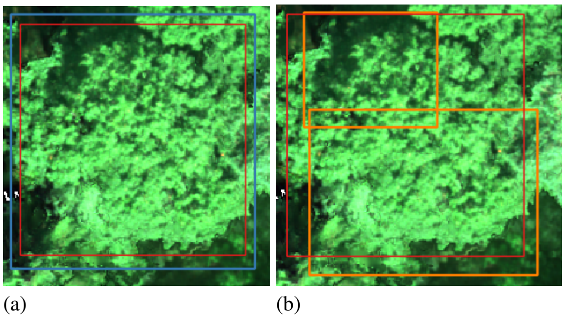


Figure 6. Sample area from mixed plot #2 showing over-segmentation. Reference bounding boxes are in red. a) Detected bounding boxes for DETR are in blue. b) Detected bounding boxes for YOLOv4 are in orange.

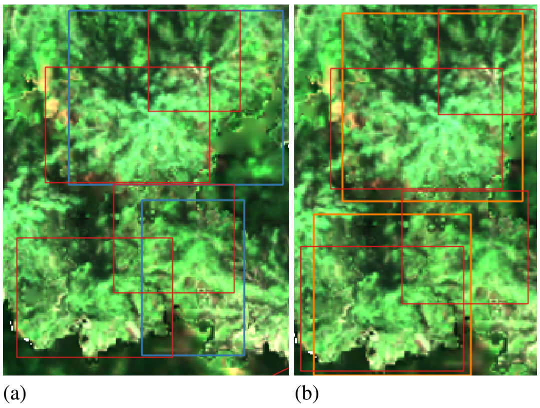


Figure 7. Sample area from coniferous plot #1 showing detection issues for small closeby trees. Reference bounding boxes are in red. a) Detected bounding boxes for DETR are in blue. b) Detected bounding boxes for YOLOv4 are in orange.

7. CONCLUSIONS AND OUTLOOK

In this study, the successful detection of individual trees using a novel transformer-based object detection method called DETR was demonstrated. When comparing DETR with the baseline method YOLOv4, we observed a significant improvement in detection accuracy. In a mixed plot, DETR achieved an improvement of more than 20% in terms of F1-score compared to YOLOv4. In a deciduous plot, a moderate increase of 4% F1-score was significant. Moreover, our experiments suggest that small trees are detected worse because of the drawbacks of DETR localising objects of reduced size.
Future experiments will focus on (i) usage of multispectral channels (e.g. NIR, NDVI, NDRE), (ii) usage of lidar-based metrics generated from a lidar flight mission conducted in the same area (e.g. DSM, lidar intensity, penetration rate), and (iii) extension of the detection method with a segmentation enabling tree crown delineation.

ACKNOWLEDGEMENTS

We thank the Research Department Life Science Systems of the Technical University Munich for the cooperation and support sensing data, especially Dr. rer. nat. Thomas Schneider for supporting us sensing data. Furthermore, we would like to thank Mr. Schlechtriem from DroneIT, who conducted the UAV flight missions. This research has been partially funded by the Federal Ministry of Education and Research (BMBF) under the contract number 13FH004IX6.

REFERENCES

Bochkovskiy, A., Wang, C., Liao, H. M., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR*, abs/2004.10934. doi.org/10.48550/arXiv.2004.10934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. *CoRR*, abs/2005.12872. doi.org/10.48550/arXiv.2005.12872.

Chen, G., Shang, Y., 2022. Transformer for Tree Counting in Aerial Images. *Remote Sensing*, 14(3). doi.org/10.3390/rs14030476.

Chen, H., Qi, Z., Shi, Z., 2021. Remote Sensing Image Change Detection With Transformers. *CoRR*, abs/2103.00208. doi.org/10.1109/TGRS.2021.3095166.

Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L., 2021. Dynamic DETR: End-to-End Object Detection With Dynamic Attention. 2988-2997. doi.org/10.1109/ICCV48922.2021.00298.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2), 303–338. doi.org/10.1007/s11263-009-0275-4.

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. *Remote Sensing*, 12(8). doi.org/10.3390/rs12081288.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385. doi.org/10.1109/CVPR.2016.90.

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. *Remote Sensing*, 12(4). doi.org/10.3390/rs12040661.

Kuhn, H. W., 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 83-97. doi.org/10.1002/nav.3800020109.

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. *International Journal of Applied Earth Observation and Geoinformation*, 42, 162-174. doi.org/10.1016/j.jag.2015.06.008.

Latifi, H., Heurich, M., 2019. Multi-Scale Remote Sensing-Assisted Forest Inventory: A Glimpse of the State-of-the-Art and Future Prospects. *Remote Sensing*, 11(11). doi.org/10.3390/rs11111260.

LeCun, Y., Huang, F. J., Bottou, L., 2004. Learning methods for generic object recognition with invariance to pose and lighting. 2, II-104 Vol.2. doi.org/10.1109/CVPR.2004.1315150.

Li, W., Guo, Q., Jakubowski, M., Kelly, M., 2012. A New Method for Segmenting Individual Trees from the Lidar Point Cloud. *Photogrammetric Engineering and Remote Sensing*, 78, 75-84. doi.org/10.14358/PERS.78.1.75.

Lin, T., Goyal, P., Girshick, R. B., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection. *CoRR*, abs/1708.02002. doi.org/10.48550/arXiv.1708.02002.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312. doi.org/10.48550/arXiv.1405.0312.

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

MicaSense, 2022. Red edge mx. `https://micasense.com/rededge-mx/`, accessed 2022-03-29.

Misra, D., 2019. Mish: A Self Regularized Non-Monotonic Neural Activation Function. *CoRR*, abs/1908.08681. doi.org/10.48550/arXiv.1908.08681.

National Ecological Observatory Network (NEON), 2022. Vegetation structure (dp1.10098.001).

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D., 2018. Image Transformer. 80, 4055–4064. doi.org/10.48550/arXiv.1802.05751.

Prangemeier, T., Reich, C., Koeppl, H., 2020. Attention-Based Transformers for Instance Segmentation of Cells in Microstructures. 700-707. doi.org/10.1109/BIBM49941.2020.9313305.

Redmon, J., Divvala, S. K., Girshick, R. B., Farhadi, A., 2015. You Only Look Once: Unified, Real-Time Object Detection. *CoRR*, abs/1506.02640. doi.org/10.48550/arXiv.1506.02640.

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

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767. doi.org/10.48550/arXiv.1804.02767.

Ren, S., He, K., Girshick, R. B., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR*, abs/1506.01497. doi.org/10.48550/arXiv.1506.01497.

Rezatofighi, S. H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. D., Savarese, S., 2019. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *CoRR*, abs/1902.09630. doi.org/10.48550/arXiv.1902.09630.

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. *Canadian Journal of Remote Sensing*, 42(5), 554-573. doi.org/10.1080/07038992.2016.1196582.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is All you Need. 30. doi.org/10.48550/arXiv.1706.03762.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., Andina, D., 2018. Deep Learning for Computer Vision: A Brief Review. *Intell. Neuroscience*, 2018. doi.org/10.1155/2018/7068349.

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. *Remote Sensing*, 11(11). doi.org/10.3390/rs11111309.

Windrim, L., Bryson, M., 2020. Detection, Segmentation, and Model Fitting of Individual Tree Stems from Airborne Laser Scanning of Forests Using Deep Learning. *Remote Sensing*, 12(9). doi.org/10.3390/rs12091469.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., Rush, A. M., 2020. Transformers: State-of-the-Art Natural Language Processing. 38–45. doi.org/10.48550/arXiv.1910.03771.

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. doi.org/10.48550/arXiv.2010.04159.