Deep learning based individual tree crown delineation from panchromatic aerial imagery

Jiaojiao Tian [1,2 *], Daniel Panangian[1], Wen Fan[2], Bastian Siegmann[3], Xiangtian Yuan[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
[2]Institute of Forest Management, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany
[3]Institute for Bio- and Geosciences, Forschungszentrum Jülich GmbH , Jülich, Germany

Keywords: Airborne Remote Sensing, ITC segmentation, colorization, delineation, deep learning, Forestry

Abstract

Accurate delineation of individual tree crowns (ITC) enables a better understanding of tree-level growth dynamics and evaluating tree vitality. In recent year, researches have introduced deep learning techniques in this field. However, the precise segmentation relies on high quality annotated dataset and test images with limited domain gaps between the training data. Under the framework of the Helmholtz project, panchromatic airborne images are captured over a mixed European forest. In this research, we adopt a UAV benchmark dataset as training data. To close the domain gaps, a deep learning based colorization step is added, for which two deep learning frameworks are compared to achieve an improved ITC delineation result in a dense forest area.

1. Introduction

With advancements in image processing techniques and the increasing availability of very high-resolution imagery, individual tree crown (ITC) delineation has gained significant attention (Holzwarth et al., 2023). ITC delineation enables the extraction of highly accurate forest inventory and analysis at the individual tree level, supporting tasks such as growth rate estimation, species classification, and health assessment. These results can assist researchers and forest managers to make more informed management decisions (Tian et al., 2020; Kempf et al., 2021).

ITC extraction from remote sensing imagery is challenging due to the similar color and texture of trees. The difficulty is especially pronounced for clumped, deciduous tree crowns, where the boundaries are sometimes hard or impossible to detect, even for human observers. Generally speaking, ITC extraction consists of two steps, tree detection and tree crown delineation (Hirschmugl et al., 2007). Beyond serving tree counting purpose, the tree detection results can also be used as seed points for tree crown delineation (Hirschmugl et al., 2007; Kempf et al., 2021).

In the last decades, numerous approaches have been developed and introduced for both ITC detection and delineation. Region-growing and watershed segmentation are the widely used traditional computer vision techniques to delineate tree crowns in images. For example, the tree crowns detected in the image were used as initial seeds for the region-growing method (Gu et al., 2020). The growing space was allocated based on Euclidean distance, and regions were expanded according to the size of the tree crowns. In comparison, seeds were set at local minimum as the treetop (Huang et al., 2018). Then, a marker-controlled watershed transformation method was performed to segment on a modified gradient size image to obtain tree crowns. Beside optical images, very high resolution digital surface model (DSM) is also an important data source for ITC delineation (Kempf et al., 2019, 2021).

In recent years, Convolutional Neural Networks (CNNs) and artificial intelligent (AI) have enabled breakthroughs in many image processing applications, including forest management (Liang et al., 2022; Zhao et al., 2023). Limited to the available annotated training data, many studies only focus on tree location/bounding box detection in less dense forest or urban areas (Zheng and Wu, 2021; Ventura et al., 2024; Luo et al., 2024; Chadwick et al., 2024). They are advanced for the transfer-learning ability and the potential for applying on large regions efficiently (Zhao et al., 2023). With more precise individual tree annotations, instance segmentation can be introduced to obtain the detailed boundaries of individual tree crowns.Yang et al. (2022) uses the classic Mask R-CNN algorithm and Google Earth images to detect canopy contours in New York’s Central Park. Although the trained model performances well for many individual trees, there are still significant erroneous extractions and omissions. With a small amount of training data, Dersch et al. (2024) proposes a semi-supervised ITC delineation approach on RGB-NIR aerial multispectral imagery and LiDAR data. Using the latest published ITC benchmark dataset (Troles et al., 2024), Fan et al. (2024) has further proved the advantages of using the Mask R-CNN model for ITC segmentation.

Most of deep learning based approaches use images from the same source or even the same region to test the trained model (Zhao et al., 2023; Luo et al., 2024; Dersch et al., 2024), which is often not the case in real-world scenarios. In addition, the use of UAV platforms is restricted in many EU countries, for both safety and privacy reasons, therefore aerial imagery is still the main data source fo for forest monitoring. In some low-cost camera systems, available data are sometimes limited to panchromatic gray-scale images. However, no tree delineation approach relies solely on panchromatic images currently, as color information is essential for effective vegetation monitoring. In some studies, near-infrared and height data are incorporate to enhance detection accuracy. This limitation poses challenges for methods that depend on spectral information and increases the domain gap between test data and existing benchmark training datasets.

To this end, in this paper we explore the possibility of involving deep-learning based colorization approach to generate synthetic red, green, blue bands, which enables the prediction the resembles real RGB images which contains the texture details and

* Corresponding author

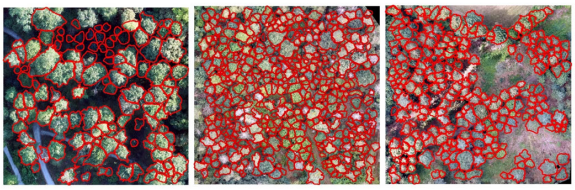

Figure 1. Examples of the BAMFORESTS dataset

resolution of original gray-scale image. Afterward, a state-of-the-art instance segmentation approach is applied to obtain high quality ITC delineations in a European natural forest.

2. Study area and data

2.1 Study area and background

Kranzberg is located in the north of Munich, Germany, which is the main test region of KROOF (Kranzberger Forest Roof Experiment) (Pretzsch et al., 2023) . It is an university test region on the special topic of forest drought stress detection. All three test regions are typical European forests, which are comprised mainly of European beech (Fagus sylvatica) and Norway spruce (Picea abies).

Within the framework of the project 3DForestSIF, airborne solar-induced fluorescence (SIF), LiDAR and panchromatic data of the Kranzberg forest were recorded in June 2023. The aim of the project is to correct the airborne SIF image data for canopy structural and illumination effects to generate information usable for the early detection of forest stress.

2.2 Aerial data

2.2.1 Panchromatic data Airborne panchromatic images of the Kranzberg forest with 80 % along and 60 % across-track overlap were recorded from 360 m above ground level with a Grasshopper U3 51S5C-C camera (Teledyne FLIR LLC, USA) on 18 June 2023 between 10:57 and 11:11 CEST. Positioning information for each image were measured with an Oxford 3052 GPS/INS unit (Oxford Technical Solutions Ltd., Oxford, UK). The single geo-tagged images were processed in Agisoft Metashape (Agisoft LLC, Russia, version 1.8.1) based on the structure from motion (SfM) technique using algorithms that identify corresponding images by feature recognition (Eltner and Sofia, 2020). As the final product of the processing, a panchromatic orthomosaic of the Kranzberg forest with a ground sampling distance of 13.8 cm was generated.

2.3 Benchmark training data

BAMFORESTS features 105 hectares of annotated, very-high-resolution UAV imagery collected using two distinct sensors from two drones. The dataset spans across four regions, including coniferous, mixed, and deciduous forests, as well as urban parks. All four forest regions are located within a radius of 20km in and around Bamberg, Germany. The original UAV images have a resolution of around 1.8 cm, with red, green and blue channels. BAMFORESTS consists of a total of 27,160 individually delineated tree crowns (ITC). Some examples of the RGB images and annotations are shown in Fig. 1. In this paper, we take the dataset with a patch size of 2048 $\times$ 2048 pixels.

3. Methodology

3.1 Example-based colorization

Example-based colorization is a task to transfer the color information from a reference image to the target grayscale image. It is a challenging task due to the ambiguity that objects can have multiple plausible color representations. Over time, various methodologies have been developed to tackle this problem, ranging from traditional statistical techniques to advanced deep learning approaches.

3.2 Histogram based

Early methods like Histogram Matching employ the pixel intensity and neighborhood statistics to find a similar pixel in the reference image and then transfer the color of the matched pixel to the target pixel. While simple and computationally efficient, this method heavily depends on the suitability of the reference image and assumes similarity in content and structure between the images. In contrast to the early methods, the deep-learning based methods are fully automatic by utilizing a large set of reference images from different categories (e.g., animal, outdoor, indoor) with various objects (e.g., tree, person, panda, car etc.).

3.3 Deep learning based colorization

Automatic colorization, which requires no reference image for color or style transfer, is a highly ill-posed problem with significant ambiguity. The same object can have various possible color representations. Cheng et al. introduced the first CNN-based colorization method (Cheng et al., 2015), demonstrating that neural networks can learn mappings from grayscale to color images. InstColor (Su et al., 2020) emphasized the importance of clear figure-ground separation and incorporated a detection model to provide object bounding boxes as priors. This concept was extended in subsequent works that utilized segmentation masks as pixel-level object semantics to guide the colorization process more precisely. More recently, some methods have aimed to restore vivid colors by leveraging the rich and diverse color priors available in pre-trained Generative Adversarial Networks (GANs) (Kim et al., 2022), capitalizing on their ability to generate high-quality, realistic images.

Transformers (Vaswani et al., 2017), initially successful in natural language processing, have been adapted for computer vision tasks, including image colorization. Transformer-based methods such as ColorFormer and DDColor both operate in the CIELAB color space, where the luminance channel (L^*) is provided by the grayscale image, and the chrominance channels (a^* and b^*) are predicted. This separation aligns with human color perception and simplifies the learning problem by allowing the models to focus solely on adding color information.

3.3.1 Colorformer Colorformer is a transformer-based image colorization framework (Ji et al., 2022) that excels in semantic consistency and color richness. It utilizes a Global-Local Hybrid (GLH) Transformer encoder with a novel Global-Local hybrid Multi-head Self-Attention (GL-MSA) mechanism to capture both global context and local details. The encoder processes grayscale images to extract multi-scale semantic representations, while the decoder upsamples features and incorporates a Color Memory (CM) module. This CM module stores semantic-color pairs derived from extensive clustering on a large dataset, providing adaptive color priors that guide the colorization process.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-885-2025 | © Author(s) 2025. CC BY 4.0 License.

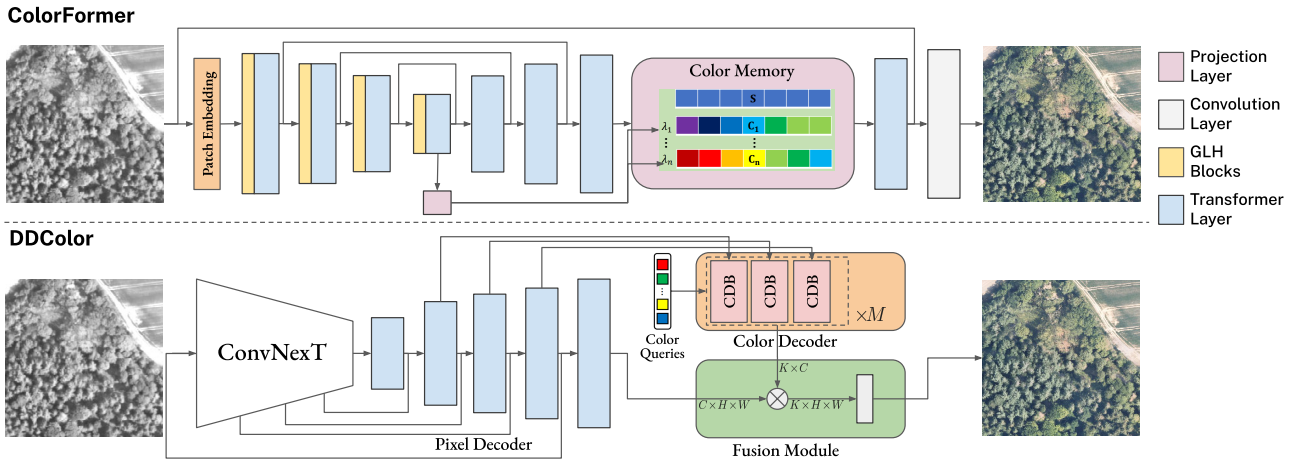

Figure 2. Architecture comparison of ColorFormer (Ji et al., 2022) and DDColor (Kang et al., 2023) for image colorization. Colorformer employs a patch embedding, GLH blocks, and transformer layers to extract features, feeding them into a Color Memory module for adaptive color priors. DDColor uses a ConvNeXT backbone, with a dual-decoder setup (Pixel and Color Decoders) that refines color embeddings via color queries. The Fusion Module combines spatial and color features to produce the final output.

The framework operates in the CIELAB color space, focusing on recovering the a^*b^* channels $x^{a^*b^*} \in \mathbb{R}^{H\times W\times 2}$ from the luminance channel $x^L \in \mathbb{R}^{H\times W\times 1}$. The encoder consists of four stacked GLH Transformer blocks, each comprising GL-MSA followed by shifted window multi-head self-attention (SW-MSA) from the Swin Transformer (Liu et al., 2021). GL-MSA extracts global features via average pooling and projects them to obtain key and value vectors, which are concatenated with local window key and value vectors to capture both local and global dependencies.

The CM module provides semantically matched color priors by storing patch-level semantic features as keys and corresponding color priors as values. Multiple color values can be assigned to one semantic, with weights determined by global semantics from the last encoder layer. Before training, keys and values are established by extracting features using a pre-trained image classification network, reducing dimensions with PCA, and clustering into $m = 512$ semantic classes using K-means. Multiple color priors for each semantic class are further clustered into $n = 4$ categories.

Training follows a generative adversarial scheme with three loss functions: (1) content loss (L_1 distance between the colorized image and ground truth), (2) perceptual loss (weighted L_1 distances between feature maps extracted by a pre-trained VGG16 (Simonyan and Zisserman, 2014)), and (3) adversarial loss from the PatchGAN discriminator (Isola et al., 2017).

The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where:

Q : query matrix
K : key matrix
V : value matrix
d_k : dimension of the keys

Multi-head attention is computed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$
$$W_i^Q \in \mathbb{R}^{d_{\text{model}}\times d_k}$$
$$W_i^K \in \mathbb{R}^{d_{\text{model}}\times d_k}$$
$$W_i^V \in \mathbb{R}^{d_{\text{model}}\times d_v}$$
$$W^O \in \mathbb{R}^{hd_v\times d_{\text{model}}}$$

3.3.2 DDColor In contrast, DDColor (Kang et al., 2023) adopts a dual-decoder architecture within an encoder-decoder framework, operating in the CIELAB color space. Given a grayscale input $x^L \in \mathbb{R}^{H\times W\times 1}$, the network predicts the missing chrominance channels $\hat{y}^{AB} \in \mathbb{R}^{H\times W\times 2}$, as illustrated in Figure 2. DDColor utilizes ConvNeXt (Liu et al., 2022) as the backbone encoder to extract high-level semantic features from grayscale images, outputting four feature maps at resolutions $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$, and $\frac{H}{32}\times\frac{W}{32}$. These features form a hierarchical representation crucial for colorization, with shortcut connections linking the encoder and decoders for maintaining spatial integrity.

The network employs two specialized decoders:

- **Pixel Decoder**: This decoder restores spatial resolution using a step-by-step upsampling process via PixelShuffle (Shi et al., 2016), enhancing the spatial detail of the image. Each upsampling layer in the pixel decoder has a shortcut connection to corresponding stages in the encoder, ensuring alignment between low- and high-resolution features. The output from the pixel decoder is an image embedding $E_i \in \mathbb{R}^{C\times H\times W}$, maintaining the input resolution.
- **Color Decoder**: The color decoder refines semantic-aware color embeddings through a series of Color Decoder Blocks (CDBs). Each CDB applies a modified transformer structure incorporating both cross-attention and multi-head self-attention to align color embeddings with visual features. A unique set of adaptive color queries, initialized as zero during training, is gradually enriched by cross-attention with the visual features at multiple scales from the pixel decoder. This multi-scale approach enhances semantic awareness and reduces color bleeding, allowing the color decoder to better identify semantic boundaries in complex

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-885-2025 | © Author(s) 2025. CC BY 4.0 License.

contexts. The final output of the color decoder, $E_c \in \mathbb{R}^{K \times C}$, represents enriched color embeddings sensitive to the semantic content of the input.

The outputs of the pixel and color decoders are combined in a lightweight fusion module to generate the final colorized image. The module uses a dot product to merge the spatial embedding E_i and the color embedding E_c, followed by a 1×1 convolution to produce the final $\hat{y}^{AB} \in \mathbb{R}^{2 \times H \times W}$ output. This colorization result is obtained by concatenating $\hat{y}^{AB}$ with the original grayscale x^L channel.

DDColor use the same adversarial training scheme and similar losses as Colorformer, including pixel-level L_1 loss, perceptual loss based on VGG16, and adversarial loss via PatchGAN. Additionally, DDColor introduces a **Colorfulness Loss** to encourage vibrant and visually appealing colors.

It is formulated as follows:

$$\mathcal{L}_{\text{col}} = 1 - \frac{\sigma_{\text{rgyb}}(\hat{y}) + 0.3 \cdot \mu_{\text{rgyb}}(\hat{y})}{100}$$

where $\sigma_{\text{rgyb}}(\cdot)$ and $\mu_{\text{rgyb}}(\cdot)$ denote the standard deviation and mean value, respectively, of the pixel cloud in the color plane, as described in (Hasler and Suesstrunk, 2003). This loss function promotes more vibrant and visually appealing colorization results by optimizing the distribution of colors in the AB channels.

3.4 ITC delineation

In this paper, Mask Region-based Convolutional Neural Network(Mask R-CNN) is the main model (He et al., 2017). The workflow is shown in Figure 3. ConvNeXt V2 is used as the backbone for individual tree segmentation, which is responsible for extracting multi-scale feature maps from the input images (Woo et al., 2023). A series of convolutional neural networks (ConvNets) dubbed ConvNeXt. ConvNeXt V2 is a new version based on that. These feature maps are then passed to the Feature Pyramid Network (FPN) to support object detection and segmentation tasks. In this experiment, the output layer of ConvNeXt V2 outputs 4 feature maps of different levels, which are fed into FPN. Through the FPN, the model is able to use feature maps with different resolutions, which helps detect both large and small objects. The inference results were filtered based on the Euclidean distance between the tree crown center and the sub-image center to remove tree crowns.

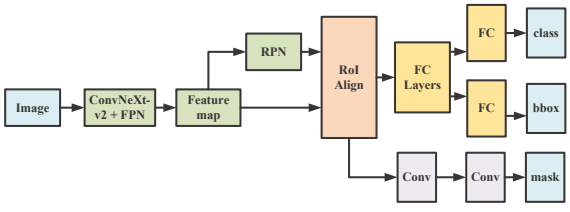

Figure 3. The workflow for ITC delineation.

4. Experiment

4.1 Colorization

Most state-of-the-art colorization models are initially trained on large-scale natural image datasets like ImageNet, which feature a diverse range of everyday scenes captured at ground level. In contrast, aerial remote sensing images are acquired from airborne or satellite platforms which produce a top-down perspective that significantly alters scale, texture, and spatial context. Additionally, these images differ in spectral characteristics because they are captured using sensors optimized for specific remote sensing applications, resulting in variations in contrast, resolution, and noise levels.

Fine-tuning these models on naturally colored aerial images (such as those available for forested areas) allows the models to adapt its learned representations to the domain-specific cues present in remote sensing data. Luckily, this task does not require any manual annotation, and abundant colored aerial images for forest are available. To train the model, 20000 patches of images 256×256 captured in Kranzberg forest are selected for training. The model is initialized from the ImageNet trained weights for fast convergence. We trained for 20000 iterations on 4 Nvidia RTX 2080 Ti GPUS with a batch size of 4 per GPU. Learning rate is set to 0.0001 initially and decayed at 8000, 12000 and 16000 iteration with gamma set to 0.5. For inference, the test image is tiled into 256×256 patches with 50% overlap. After inference, the patches are mosaicked back to the final raster. To alleviate border artefacts, we employed a weight matrix on each patch, where the weights are smaller when closer to the patch border and higher when closer to the center. The quality of the colorization was evaluated through qualitative visual inspection by domain experts. The experts assessed the visual realism focusing on natural color distributions, edge fidelity, and the absence of artifacts.

4.2 ITC delineation

The method is implemented using MMDetection 3.0 (Chen et al., 2019) and trained on an Nvidia GeForce RTX 3080 GPU. The base version of ConvNeXt V2 was used in this process. The pixel size of the training data is 2048 $\times$2048. As there are not supposed to be a significant difference between the inference images and the training data, the Kranzberg forest images are resampled to a resolution of 0.2m, with each sub-image having a pixel size of 2048x2048. The sub-images overlap by 50% horizontally and vertically. In addition, the training process consists of 50 epochs, with a 50% minimum Intersection over Union (IoU) threshold maintained for both bounding box predictions and ground truth boxes. The optimizer uses the AdamW optimizer with a learning rate of 0.0001 and weight decay of 0.05.

4.3 Results and Discussion

4.3.1 Comparison of colorization approaches We compared the performance of three colorization methods—Histogram Matching, DDColor, and ColorFormer—across a variety of landscapes, including forested areas, agricultural fields, and urban area. The results are shown in fig. 5. Histogram Matching struggled with realistic color representation, particularly in urban or semi-urban scenes. For instance, in forested areas, uniform dark green tones flatten the scene and fail to capture the dynamic nature of forest canopies. In the rural town environment, most of the rooftops are incorrectly rendered in green, blending with the surrounding vegetation and severely reducing the contrast needed to distinguish buildings from their environment. Moreover, Histogram Matching produced the least texture variation, resulting in relatively flat representations across all scenes. Trees appeared uniform and lacked the shading necessary to depict depth, which was especially noticeable in the forested areas.

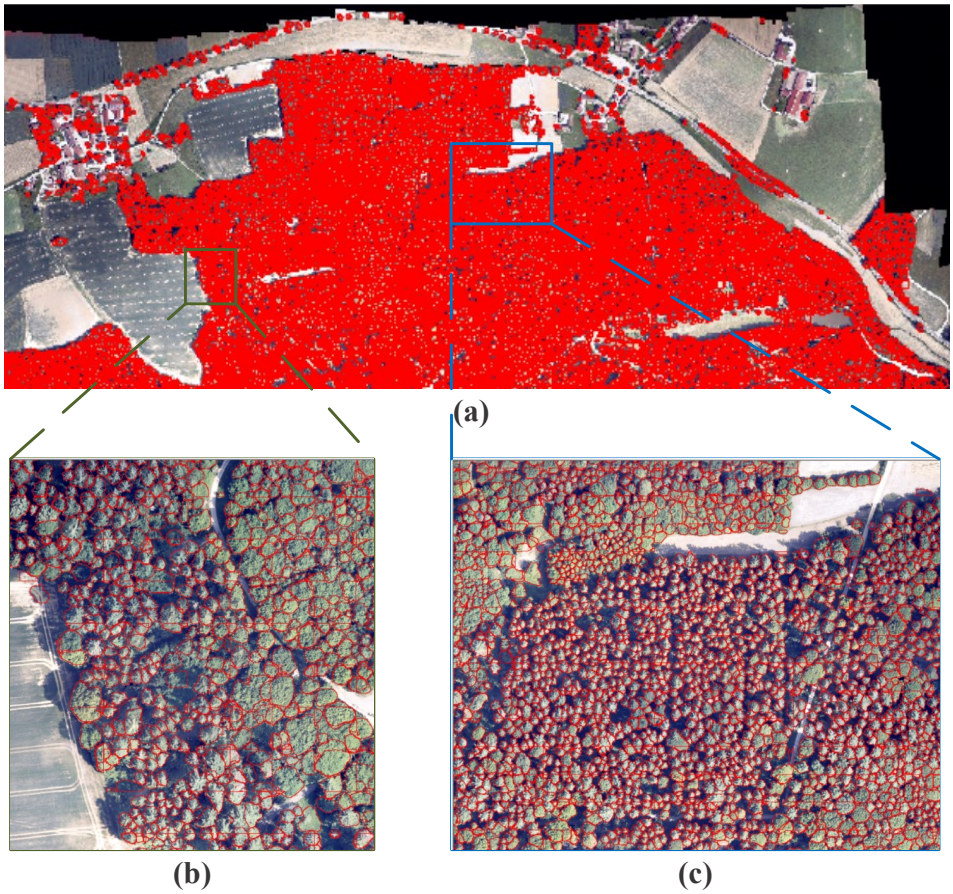

Figure 4. ITC results for (a) Overall results, (b) and (c) are parts of (a).

The deep learning methods, DDColor and ColorFormer, both improved upon Histogram Matching by providing better color variation and texture overall. In the other hand, both still had issues with consistently coloring rooftops. Green tones from nearby vegetation sometimes spilled onto roofs, and some roof parts were unevenly colored. When comparing the two models, DDColor provided better color variation and texture but introduced a yellowish-green tint in forested and agricultural areas. Oversaturation made scenes appear unnatural and hindered fine-scale tasks like identifying individual tree crowns. In contrast, ColorFormer consistently outperformed the other two methods in terms of visual realism. The method avoided the oversaturation and yellowish tones found in DDColor and delivered balanced and natural colors across all scenes. Therefore we selected ColorFormer for the colorization due to its performance on forested environments.

4.3.2 ITC segmentation The results of instance segmentation in the experiments are shown in Figure 4. A qualitative assessment of the results in Figure 4 shows that the algorithm has high-quality visual performance, with ConvNeXt V2 effectively distinguishing between tree crowns of different scales. Figures 4(b) and 4(c) show that the majority of tree crowns can be effectively segmented. The contours of the crowns in figure 4(c) correspond well with the manual predictions and show little under-segmentation. However, in Figure 4(b), some crowns show over-segmentation. This is due to the initial stage of the inference process, where the images at the edges were over-segmented. Therefore, Figure 4(a) is composed of many sub-images, even after the removal of duplicate tree crowns, some tree crowns located at the image edges were split into several smaller crowns. These results are still retained and require further exploration.

Kranzberg is featured as a mixed natural forest with both broadleaf and coniferous trees. For larger crowns, segmentation performance is varied between tree species. As shown in Figure 6, the algorithm performs well in segmenting coniferous trees. However, there is some over-segmentation at the edges of sub-images, most likely because the sub-images themselves divide the tree crowns into multiple parts. The model also performs well for broadleaved trees, but there are some challenges and the performance is varied between tree species. The tree crowns are contiguous, making it difficult to judge directly how to separate overlapping crowns from the image. As a result, some tree crowns may not be correctly segmented, leading to under-segmentation in Figure 6. Therefore, future work will focus on improving the performance of the model to deal with both over-segmentation and under-segmentation for different tree species.

In the absence of ground truth data for Kranzberg, only visual qualitative assessment is provide. The ITC delineation on the grey scale image and the colorized image are compared and shown in Figure 7. The differences in image values between the individual bands are small, while the three-band information is more abundant. In grey images, tree boundaries are not clear and tree centres are difficult to identify, resulting in fewer detec-

Figure 5. Comparison of colorization methods on different scenes. The first column shows the panchromatic image input, followed by the colorized outputs.

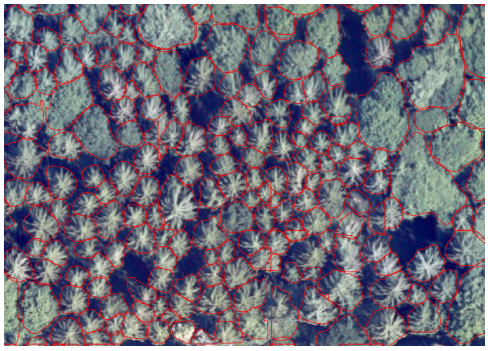

Figure 6. Part area of the ITC delineation result.

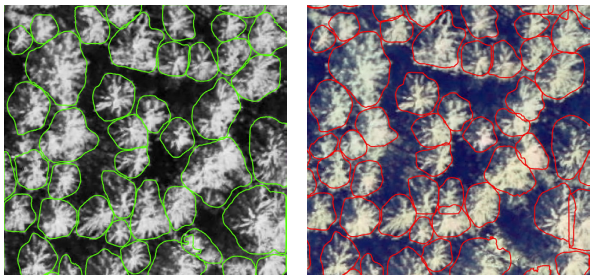

Figure 7. Comparison of ITC delineation results in grey value images and colorized images.

tions. In the colorized images, the differences between different tree crowns are obvious. As shown in Figure 7 (b), the centers and distributions of conifer crowns are clear. This helps to better identify the existing of individual trees, with fewer false negatives.

5. Conclusion

Deep learning models have great potential to automate the forest inventory measurement, especially when more pubic benchmark data are available. In this study we have successfully performed the ITC delineation using panchromatic images. In order to bridge the domain gap between our panchromatic data and RGB images and to enrich the available spectral features, a colorization step is introduced to generate artificial RGB images. Compared to the traditional histogram-based approach, deep learning-based approaches are able to generate an image with rich and close to real spectral features. Tree boundaries can be better represented, further improving ITC delineation results. In our experiment, the ITC delineation model trained on UAV data has performed well on the aerial dataset. Unfortunately we can not provide a numerical evaluation due to the lack of ground truth data. More details on tree-level based analyses will be performed in the further studies.

References

Chadwick, A. J., Coops, N. C., Bater, C. W., Martens, L. A., White, B., 2024. Transferability of a mask R–CNN model for the delineation and classification of two species of regenerating tree crowns to untrained sites. *Science of Remote Sensing*, 9, 100109.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C.,

Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., Lin, D., 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*.

Cheng, Z., Yang, Q., Sheng, B., 2015. Deep colorization. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, IEEE Computer Society, USA, 415–423.

Dersch, S., Schöttl, A., Krzystek, P., Heurich, M., 2024. Semi-supervised multi-class tree crown delineation using aerial multispectral imagery and lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216, 154–167.

Eltner, A., Sofia, G., 2020. Structure from motion photogrammetric technique. *Developments in Earth surface processes*, 23, Elsevier, 1–24.

Fan, W., Tian, J., Troles, J., Döllerer, M., Kindu, M., Knoke, T., 2024. Comparing Deep Learning and MCWST Approaches for Individual Tree Crown Segmentation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 67–73.

Gu, J., Grybas, H., Congalton, R. G., 2020. Individual tree crown delineation from UAS imagery based on region growing and growth space considerations. *Remote Sensing*, 12(15), 2363.

Hasler, D., Suesstrunk, S., 2003. Measuring Colourfulness in Natural Images. *Proceedings of SPIE - The International Society for Optical Engineering*, 5007, 87-95.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hirschmugl, M., Ofner, M., Raggam, J., Schardt, M., 2007. Single tree detection in very high resolution remote sensing data. *Remote Sensing of Environment*, 110(4), 533–544.

Holzwarth, S., Thonfeld, F., Kacic, P., Abdullahi, S., Asam, S., Coleman, K., Eisfelder, C., Gessner, U., Huth, J., Kraus, T. et al., 2023. Earth-observation-based monitoring of forests in Germany—recent progress and research frontiers: a review. *Remote Sensing*, 15(17), 4234.

Huang, H., Li, X., Chen, C., 2018. Individual Tree Crown Detection and Delineation From Very-High-Resolution UAV Images Based on Bias Field and Marker-Controlled Watershed Segmentation Algorithms. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(7), 2253-2262.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Ji, X., Jiang, B., Luo, D., Tao, G., Chu, W., Xie, Z., Wang, C., Tai, Y., 2022. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. *European Conference on Computer Vision*, Springer, 20–36.

Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X., 2023. Ddcolor: Towards photo-realistic image colorization via dual decoders. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 328–338.

Kempf, C., Tian, J., Kurz, F., d'Angelo, P., Reinartz, P., 2019. Local Versus Global Variational Approaches to Enhance Watershed Transformation Based Individual Tree Crown Segmentation of Digital Surface Models from 3k Optical Imagery. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*, 42(2), 43–50.

Kempf, C., Tian, J., Kurz, F., D'Angelo, P., Schneider, T., Reinartz, P., 2021. Oblique view individual tree crown delineation. *International Journal of Applied Earth Observation and Geoinformation*, 99, 102314.

Kim, G., Kang, K., Kim, S., Lee, H., Kim, S., Kim, J., Baek, S.-H., Cho, S., 2022. Bigcolor: Colorization using a generative color prior for natural images. *European Conference on Computer Vision (ECCV)*.

Liang, X., Kukko, A., Balenović, I., Saarinen, N., Junttila, S., Kankare, V., Holopainen, M., Mokroš, M., Surovỳ, P., Kaartinen, H. et al., 2022. Close-Range Remote Sensing of Forests: The state of the art, challenges, and opportunities for systems and data acquisitions. *IEEE geoscience and remote sensing magazine*, 10(3), 32–71.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

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

Luo, T., Gao, W., Belotserkovsky, A., Nedzved, A., Deng, W., Ye, Q., Fu, L., Chen, Q., Ma, W., Xu, S., 2024. VrsNet-density map prediction network for individual tree detection and counting from UAV images. *International Journal of Applied Earth Observation and Geoinformation*, 131, 103923.

Pretzsch, H., Ahmed, S., Rötzer, T., Schmied, G., Hilmers, T., 2023. Structural and compositional acclimation of forests to extended drought: results of the KROOF throughfall exclusion experiment in Norway spruce and European beech. *Trees*, 37(5), 1443–1463.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1874–1883.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Su, J.-W., Chu, H.-K., Huang, J.-B., 2020. Instance-aware image colorization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, J., Schneider, T., Kempf, C., Xia, Y., Lusseau, M., Hill, J., Jachmann, E., Reinartz, P., 2020. Early detection of forest drought stress with very high resolution stereo and hyperspectral imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 781–787.

Troles, J., Schmid, U., Fan, W., Tian, J., 2024. BAMFORESTS: Bamberg Benchmark Forest Dataset of Individual Tree Crowns in Very-High-Resolution UAV Images. *Remote Sensing*, 16(11), 1935.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., Polosukhin, I., 2017. Attention is all you need. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems*, 30, Curran Associates, Inc.

Ventura, J., Pawlak, C., Honsberger, M., Gonsalves, C., Rice, J., Love, N. L., Han, S., Nguyen, V., Sugano, K., Doremus, J. et al., 2024. Individual tree detection in large-scale urban environments using high-resolution multispectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, 130, 103848.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16133–16142.

Yang, M., Mou, Y., Liu, S., Meng, Y., Liu, Z., Li, P., Xiang, W., Zhou, X., Peng, C., 2022. Detecting and mapping tree crowns based on convolutional neural network and Google Earth images. *International Journal of Applied Earth Observation and Geoinformation*, 108, 102764. https://www.sciencedirect.com/science/article/pii/S0303243422000903.

Zhao, H., Morgenroth, J., Pearse, G., Schindler, J., 2023. A systematic review of individual tree crown detection and delineation with convolutional neural networks (CNN). *Current Forestry Reports*, 9(3), 149–170.

Zheng, Y., Wu, G., 2021. Single shot multibox detector for urban plantation single tree detection and location with high-resolution remote sensing imagery. *Frontiers in Environmental Science*, 9, 755587.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-885-2025 | © Author(s) 2025. CC BY 4.0 License.