# Tree Species Classification Using Majority Voting Approach on UAV Raw Images

Tomohiro Mizoguchi[1], Daisuke Tsukano[2], Hideki Ogawa[3]

[1]Department of Informatics and Data Science, Faculty of Engineering, Sanyo-Onoda City University, Sanyo-Onoda, Yamaguchi, Japan - mizoguchi@rs.socu.ac.jp
[2]Owada Survey & Design, Futaba-gun, Fukushima, Japan - d_tsukano@geo999.com
[3]Fukushima Prefectural Forestry Research Center, Koriyama City, Fukushima, Japan - ogawa_hideki_01@pref.fukushima.lg.jp

**Keywords:** Tree Species, UAV, Deep Learning, Raw Image, Majority voting.

**Abstract**

In recent years, the widespread availability of relatively low-cost cameras and unmanned aerial vehicles (UAVs) has made it easier to acquire high-resolution images of forests rich in texture information. In forest resource surveys, a common approach for tree species classification involves applying Structure from Motion (SfM) to the collected raw image set and using the resulting orthophotos for classification through deep learning techniques. This study proposes a novel tree species classification framework that aims to improve accuracy by utilizing UAV-acquired raw images, which are of high quality, can be captured in large volumes, and include diverse viewing angles. In the proposed method, tree species classification is first performed on each aerial image using a convolutional neural network (CNN) and a sliding window approach. Next, SfM processing is applied to the image set to generate a 3D point cloud and orthophotos. The classification results from the aerial images are then projected onto the point cloud, and finally, these projected results are mapped onto the orthophoto and aggregated to derive the final species classification by majority voting approach. The effectiveness of the proposed method is validated through experiments targeting four representative tree types in Japan: *Cryptomeria japonica* (Japanese cedar), *Chamaecyparis obtusa* (Japanese cypress), *Pinus densiflora* (Japanese red pine), and broadleaf trees.

## 1. Introduction

### 1.1 Background

Accurate tree species classification is a fundamental and critically important process for the sustainable management and efficient utilization of forest resources. Traditionally, tree species classification has been conducted through field surveys by experts or visual interpretation of aerial photographs. However, field surveys are often associated with significant time, cost, and labor, posing substantial operational burdens. Moreover, visual interpretation is highly dependent on factors such as image quality and resolution, as well as the experience and skill of the analyst, which can lead to inconsistencies in results (Lee, 2023).

In recent years, unmanned aerial vehicles (UAVs) equipped with consumer-grade cameras have become widely used in forest surveys. UAVs can fly at low altitudes at relatively slow speeds, allowing for efficient capture of high-resolution images rich in texture information over large areas spanning several hectares. Compared with other platforms, UAVs offer greater flexibility and fewer constraints in terms of time and location for data acquisition. As a result, UAVs are becoming essential tools for tree species classification, which requires capturing fine-scale features of tree crowns.

In addition, the rapid advancement of deep learning technologies in recent years has enabled high performance across a wide range of image processing tasks. One of the key advantages of deep learning lies in its powerful feature extraction capabilities, which allow it to effectively capture species-specific characteristics, even in fine-grained classification problems such as tree species identification. Consequently, many studies have reported higher classification accuracy compared to conventional methods. Thus, the approach of acquiring images from UAV-mounted consumer cameras, constructing large-scale training datasets, and applying deep learning for tree species classification is increasingly being established as a promising and practical solution.

### 1.2 Related Works

In recent years, tree species classification using convolutional neural networks (CNNs) applied to UAV imagery has emerged as a prominent research topic, with various methods being proposed. One of the most fundamental approaches involves applying Structure from Motion (SfM) processing to RGB images acquired at a specific time to generate orthophotos, followed by tree species classification using deep learning techniques. Classification methods based on deep learning can generally be categorized into two groups: semantic segmentation and classification. Semantic segmentation (SS) methods, such as U-Net and DeepLabv3, aim to assign tree species labels to each pixel of the orthophoto with high accuracy (Schiefer, 2020; Popp, 2023). In contrast, classification approaches typically involve segmenting the image into individual tree crowns in advance, and then applying a CNN model (e.g., ResNet) to classify each crown, enabling tree-level species identification. Segmentation methods used in this context include the watershed algorithm (Natesan, 2019) and auxiliary segmentation techniques based on LiDAR point clouds (Ma, 2024). Additionally, a simpler approach has been reported in which the image is divided into small, regularly spaced regions, and classification is performed for each region independently (Huang, 2023).

The introduction of deep learning has enabled higher classification accuracy compared to conventional methods; however, various efforts are being made to further improve performance. For example, extensive research has been conducted on the use of multispectral and hyperspectral imagery. While standard digital cameras capture three spectral bands, red, green, and blue, multispectral cameras can acquire 4 to 8 bands including near-infrared (NIR), and hyperspectral cameras can capture more than 100 bands, including NIR and shortwave infrared (SWIR) (Onishi, 2022). By incorporating these non-visible spectral bands, it becomes possible to evaluate the characteristics of tree crowns in greater detail, leading to improved classification accuracy. In particular, studies have reported that integrating vegetation indices derived from NIR information as input features enhances classification

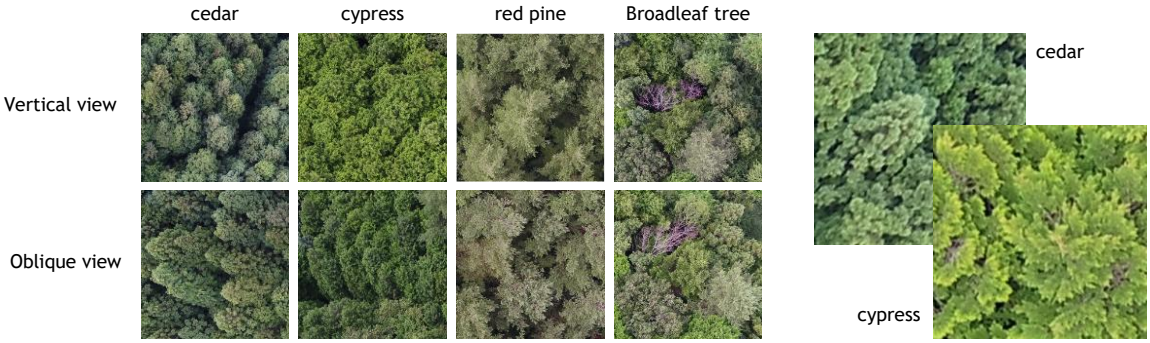


Figure 1. Visual features captured from raw aerial images.

performance (Lee, 2023; Ma, 2024). However, a key challenge is the high cost associated with acquiring and processing high-resolution hyperspectral and multispectral images.

The use of UAV-mounted LiDAR has also been reported in recent studies. In recent years, relatively low-cost LiDAR sensors that can be mounted on UAVs have become available, and some research has utilized canopy height models (CHMs) generated from the acquired 3D point clouds for tree species classification (Lee, 2023). Because point cloud data allows for capturing the structural characteristics of trees, it contributes to improving classification accuracy. Furthermore, there are studies that have achieved high classification performance by integrating features extracted from point clouds, such as height, crown shape, and reflectance intensity, into machine learning models (Zhong, 2022; Ma, 2024). However, UAV LiDAR still involves additional costs, and since it requires a separate flight from image acquisition, its application in the field remains subject to time-related constraints.

In addition, analysis methods using multi-temporal imagery have been proposed; however, issues such as extended acquisition periods and increased costs have been pointed out (Natesan, 2019; Veras, 2022; Avtar, 2024).

Given the above background, there is a continued strong demand for the development of a highly accurate tree species classification method that relies solely on single-date imagery captured using low-cost, easy-to-handle RGB cameras.

### 1.3 Key observation

This section compares the characteristics of orthophotos and raw aerial images. First, compared to orthophotos, raw images are significantly more numerous, resulting in a much larger volume of information. Moreover, as shown in Figure 1, raw images offer greater diversity in terms of viewing angles. Specifically, they include not only nadir (directly overhead) views but also oblique views taken from various angles, providing a wide range of perspectives. In particular, oblique images capture the three-dimensional structure of tree crowns more clearly. Due to their high resolution, these images also contain features, such as the arrangement of leaves on the sides of tree crowns, that are difficult to observe from overhead views alone. Additionally, because of the complex structure and appearance of trees, the quality of orthophotos is sometimes insufficient. From the perspective of image quality as well, raw images have been reported to offer superior visual information (Avtar, 2024).

Based on the above, raw aerial images are superior to orthophotos in terms of information content, diversity of viewing angles, and image quality. By effectively leveraging these advantages, higher-accuracy tree species classification can be expected compared to conventional methods that rely solely on orthophotos.

### 1.4 Purpose

In this study, we propose a novel tree species classification framework that generates a species map by performing classification on individual aerial images and aggregating the results onto a single orthophoto via a 3D point cloud. The objective is to experimentally validate the effectiveness of the proposed method.

The proposed framework consists of the following four steps, as illustrated in Figure 2. In Step 1, tree species classification is performed on multiple aerial images using a sliding window approach combined with deep learning. In Step 2, a point cloud and an orthophoto are generated from the aerial images using Structure from Motion (SfM) and Multi-View Stereo (MVS) processing, and camera parameters, including the position and orientation of each image, are estimated. In Step 3, the tree species scores computed from the aerial images are assigned to the corresponding 3D points based on image geometry. In Step 4, the species scores associated with each point are orthogonally projected onto the orthophoto, and the scores are aggregated for each pixel. The tree species label with the highest score is then assigned to each pixel, resulting in the final species map.

As shown in Figure 1, the classification targets in this study include four representative tree types in Japan: *Cryptomeria japonica* (Japanese cedar), *Chamaecyparis obtusa* (Japanese cypress), *Pinus densiflora* (Japanese red pine), and broadleaf trees.

### 1.5 Test Site

In this study, aerial images acquired in the Kawauchi Experimental Forest, located in Kawauchi Village, Fukushima Prefecture, were used. Image acquisition was conducted using a drone equipped with a camera, flying at a speed of approximately 5.5 m/s. Images were captured at a frequency of one frame every two seconds. The UAV used was the DJI Matrice 210 RTK, as shown in Figure 3, and the camera was the DJI Zenmuse X5S. The data were collected in May 2022 under mostly cloudy weather conditions. The acquired images had a resolution of 5,280 × 3,956 pixels, with an overlap and sidelap rate of 80% and 70%, respectively.

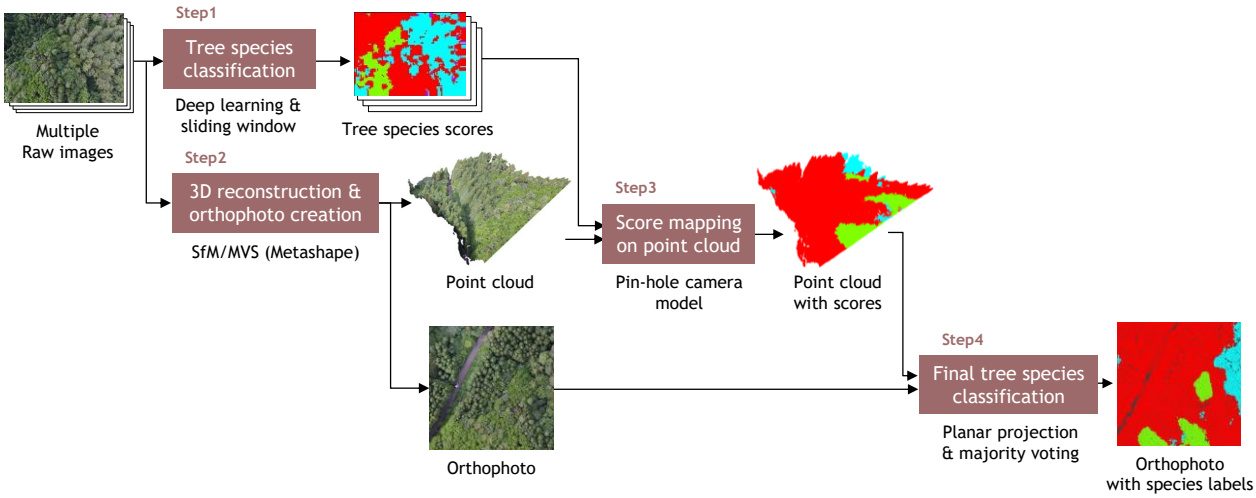


Figure 2. Overview of our proposed framework.

Figure 3. UAV used in this study.

## 2. Proposed Method

### 2.1 Tree Species Classification using Deep Learning

In the proposed method, tree species classification is first performed on raw aerial images using deep learning. The deep learning model employed is VGG16 (Simonyan, 2015), a type of convolutional neural network (CNN). A CNN is a neural network architecture composed of multiple functional layers, such as convolutional and pooling layers. Among CNN models, VGG16 is widely used as a classical model due to its relatively simple structure and stable classification performance.

Tree species classification is performed using a sliding window approach. Specifically, a target region of size $n \times n$ and a surrounding context region of size $N \times N$, centered on the target region, are defined. During classification, the window is slid across the image at intervals of $n$ pixels. At each position, the corresponding $N \times N$ context region is extracted and used as input for deep learning-based classification. The resulting prediction is then assigned to the associated $n \times n$ target region. In this study, the parameters were set to $N = 224$ and $n = 56$. As a result of this process, each pixel is assigned a set of class probabilities for tree species, obtained through the softmax function. In general, each pixel is classified into the class with the highest score. In this paper, we refer to these probabilities as "scores." Examples of a raw aerial image and the corresponding tree species classification result are shown in Figures 4(a) and 4(e), respectively. In Figure 4(e), the colors are assigned based on the label of the class with the highest score.

### 2.2 3D Reconstruction and Orthophoto Creation

For 3D reconstruction and orthophoto generation, the photogrammetry software Agisoft Metashape was used. By inputting multiple aerial images, the software automatically estimates camera parameters, including position and orientation at the time of capture, and performs 3D point cloud and orthophoto creation.

Figure 4(b) shows the generated point cloud, while Figure 4(c) presents the resulting orthophoto. The point cloud was uniformly downsampled at 2 cm spatial intervals, and the spatial resolution of the orthophoto was set to 2 cm/pixel. Figure 4(d) shows the height map of the study area, which has an elevation difference of up to approximately 74 meters.

### 2.3 Mapping of Classification Score on Point Cloud

This section describes the method for mapping tree species classification results obtained from aerial images onto a 3D point cloud. The mapping process is carried out based on the principles of the pinhole camera model (Saovana, 2021; Hartley, 2023).

Ideally, one possible approach involves generating a mesh from the point cloud, identifying the corresponding point on the mesh for each pixel, assigning the tree species score to that point, and finally projecting the score-assigned points onto the orthophoto for aggregation. However, this method requires constructing a high-density mesh to ensure accurate score assignment, resulting in a large data volume. Moreover, searching for corresponding points on the mesh entails a high computational cost.

To address these issues, this study adopts a reverse approach: for each 3D measurement point, the corresponding pixel on each aerial image is identified, and the tree species scores assigned to that pixel are then transferred to the 3D point. This approach allows for reduced data volume and more efficient processing.

The specific procedure leading to the final projection onto the orthophoto begins with computing the corresponding pixel for each point in the point cloud on each aerial image. For this purpose, the coordinates of the 3D points are transformed

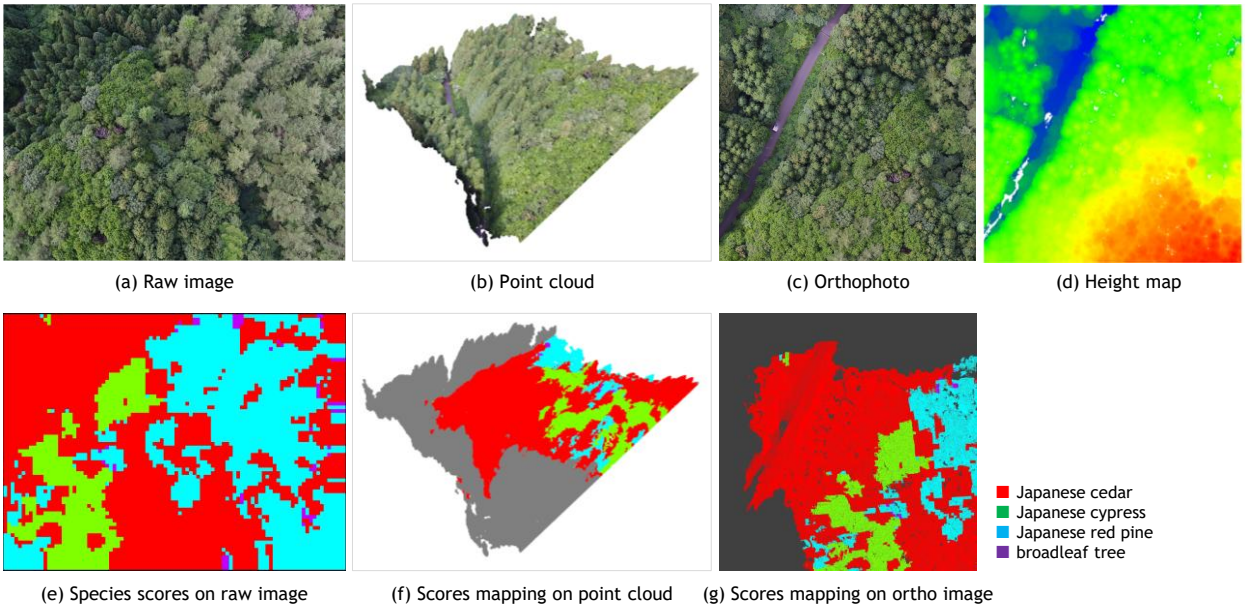


Figure 4. Tree species classification process based on proposed method.

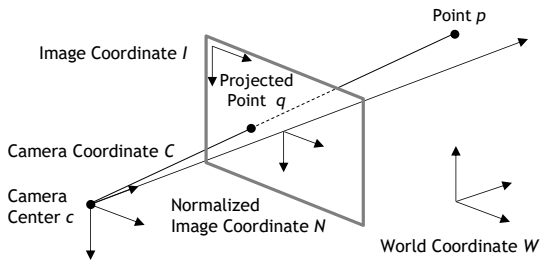


Figure 5. Coordinate transformation.

according to the following steps. This section focuses on the processing steps for a single aerial image. Subsequently, we describe how the classification results from multiple images are aggregated on the point cloud and finally projected and visualized on the orthophoto. An overview of the entire process is shown in Figure 5.

### 2.3.1 Mapping of Classification Scores of a Single Image

The coordinates $\mathbf{p}_i^W$ of each point in the point cloud $P^W$, defined in the reference world coordinate system $W$, are transformed into coordinates $\mathbf{p}_i^C = (x_i^C, y_i^C, z_i^C)$ in the camera coordinate system $C$, where the camera position $\mathbf{c} = (c_x, c_y, c_z)$ at the time of image capture is set as the origin. This transformation is performed using a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, as expressed in Equation (1).

$$\mathbf{p}_i^C = \mathbf{R}\mathbf{p}_i^W + \mathbf{t} \tag{1}$$

The rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$, and camera position $\mathbf{c}$ are all exported as output files from the SfM processing performed using Metashape.

Next, the 3D coordinates $\mathbf{p}_i^C$ in the camera coordinate system $C$ are projected onto the normalized image coordinate system $N$ as 2D coordinates $\mathbf{q}_i^N = (u_i^N, v_i^N)$. The coordinate system $N$ can be considered to lie on a virtual image plane placed at $z = 1$ in the camera coordinate system $C$. The transformation from to $\mathbf{q}_i^N$ is performed according to Equation (2).

$$(u_i^N, v_i^N) = \left(\frac{x_i^C - c_x}{z_i^C - c_z}, \frac{y_i^C - c_y}{z_i^C - c_z}\right) \tag{2}$$

Finally, the projected point $\mathbf{q}_i^N$, expressed in the normalized image coordinate system $N$, is transformed into the coordinate $\mathbf{q}_i^I$ in the intrinsic image coordinate system $I$, which is specific to the camera used for image acquisition. This transformation is represented by Equation (3).

$$(u_i^I, v_i^I) = \left(\frac{f u_i^N}{s} + u_0, \frac{f v_i^N}{s} + v_0\right) \tag{3}$$

Here, $f$ denotes the focal length, and $s$ represents the image size. The parameters $u_0$ and $v_0$ correspond to the coordinates of the intersection point between the camera's optical axis and the image plane, commonly referred to as the principal point (typically located at the center of the image).

Through the above process, the corresponding pixel on the image can be determined for each point $p_i$ in the point cloud. This enables the assignment of the tree species classification results obtained from the aerial images in Section 2.1 to the 3D point cloud. Figure 4(f) shows the result of assigning the classification outputs from Figure 4(e) onto the point cloud.

### 2.3.2 Computation of Score of Multiple Images

Since multiple aerial images are available, each measurement point may have up to as many tree species scores as the number of images. The final label is determined through weighted majority voting based on these class scores. In a simple majority voting scheme, the class with the highest number of votes is assigned as the final label for the point; however, cases may occur where multiple classes receive the same number of votes, resulting in a tie (Kokkinos, 2014; Misra, 2020). To avoid such ambiguity, this study employs a method in which the outputs of the softmax function are accumulated with weights, and the final label is determined based on the aggregated scores.

Here, for each measurement point $p_i$, the corresponding pixel $q_{j,c(i)}$ is identified from each of the aerial image $I_j$. The score corresponding to class $k$, obtained from the softmax output of the

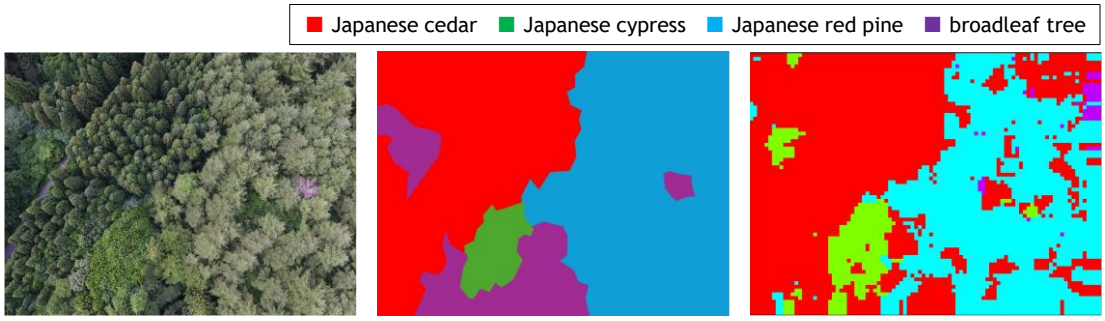


Figure 6. Tree species classification results on aerial image.

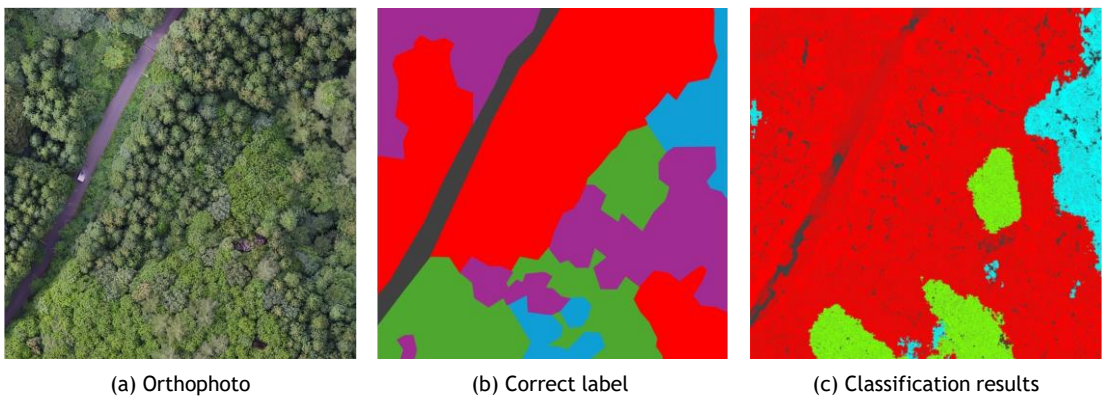


Figure 7. Assignment of tree species classification results to orthophoto.

CNN at this pixel, is denoted as $r_{j,c(i),k}$. The weighted sum of these scores is then calculated according to Equation (4).

$$R_{i,k} = \sum_{j=1}^{M} w_{j,c(i)} r_{j,c(i),k} \quad (4)$$

Here, $w_{j,c(i)}$ represents a weight, which is being considered for calculation based on factors such as the distance from the image capture location to each point and a shadow region index. Pixels that are located far from the capture position, making high-resolution classification difficult, or that fall within shadowed areas are assigned lower weights to reduce their influence on the final classification result. Currently, all weights are set uniformly as $w_{j,c(i)} = 1$.

### 2.4 Final Tree Species Classification by Majority Voting

Next, the classification results assigned to the point cloud are aggregated onto the orthophoto through orthogonal projection to obtain the final classification output. The xy-plane is divided into a grid, and the point cloud is projected onto this plane. For each projected point, the corresponding grid cell is identified, and the scores of the points within each grid cell are summed. Figure 4(g) shows the result of projecting the classification output from the single aerial image in Figure 4(e) onto the orthophoto.

Finally, the tree species label corresponding to the class with the highest aggregated score in each grid cell is assigned, and the result is output as the final tree species map. In this study, the grid size was set to 4 cm, taking point cloud density into account, in order to minimize the occurrence of empty grid cells with no included points.

## 3. Experimental Results

### 3.1 Data Set

Under the guidance of domain experts, tree species were manually labelled through visual interpretation. Based on the labelled data, regions corresponding to each species were extracted from the aerial images and then divided into patches of 224 × 224 pixels. For the dataset, 2,000 image patches were prepared for each of the four categories: *Cryptomeria japonica* (Japanese cedar), *Chamaecyparis obtusa* (Japanese cypress), *Pinus densiflora* (Japanese red pine), and broadleaf trees. Images that spanned across boundaries of different tree species were excluded, and only those containing a single species were selected through manual inspection. Although both the training and validation data were collected from the same area on the same day, they were distributed across different forest stands to ensure separation between the two sets.

### 3.2 Accuracy Evaluation on Single Raw Aerial Image

First, the classification performance on raw aerial images was qualitatively evaluated through visual inspection. While good results were obtained for some images, others exhibited noticeable misclassifications. For example, as shown in Figure 6, misclassification frequently occurred in certain areas on the right side of the image where *Pinus densiflora* (Japanese red pine) was densely distributed, and in the lower central area where *Chamaecyparis obtusa* (Japanese cypress) was present, with these regions often being misidentified as *Cryptomeria japonica* (Japanese cedar). These errors were commonly observed in boundary regions between adjacent trees. This area is located at a high elevation and was captured at a relatively short distance, resulting in large tree crowns appearing in the images.

Consequently, individual trees did not fully fit within the 224 × 224 window used in Step 1, leading to inaccurate classification. One possible countermeasure is to use multiple images with different resolutions and aggregate the classification results. Applying the same window size to lower-resolution images would allow individual trees, or at least multiple trees, to be fully contained within the window, potentially improving classification accuracy. Similar misclassification was also observed in shadowed regions. To address this, a promising approach is to pre-identify shadow areas and assign them lower weights during the voting process, thereby reducing their influence on the final classification result. Additionally, the classification accuracy for broadleaf trees tended to be lower, which is expected to improve with the inclusion of more training data.

### 3.3 Accuracy Evaluation on Orthophoto

Finally, the aggregated classification results on the orthophoto were evaluated. As shown in Figure 7, apart from areas densely populated by a single species, such as *Chamaecyparis obtusa* (Japanese cypress) and *Pinus densiflora* (Japanese red pine), the majority of regions were classified as *Cryptomeria japonica* (Japanese cedar). As mentioned earlier, misclassification into *Cryptomeria japonica* (Japanese cedar) was frequently observed in the results on individual aerial images, and this trend persisted in the final classification results on the orthophoto. Additionally, during the mapping of classification results from the raw images to the point cloud, points located behind foreground tree, thus not actually visible in the images, were nonetheless associated with visible regions due to the geometric projection process. This led to inappropriate label assignments in those areas. To address this issue, we aim to introduce a visibility check for the point cloud in future implementations.

## 4. Conclusion and Future Works

In this study, we proposed a novel framework that integrates raw UAV images with 3D point clouds and orthophotos to accurately aggregate tree species classification results. The effectiveness of the proposed method was validated through experiments targeting representative tree species.

Future work includes further improving classification accuracy through the integrated use of multi-resolution images, detecting shadow regions and estimating their influence on the final classification results, and introducing visibility analysis for the 3D point cloud. We also plan to investigate the use of semantic segmentation to enhance pixel-level classification performance.

Additionally, conducting validation experiments at multiple test sites with varying geographic conditions and vegetation compositions will be important for quantitatively evaluating the generalizability and robustness of the proposed method.

## Acknowledgements

This study was conducted as part of the commissioned research project under the Advanced Technology Development Program for Agriculture, Forestry and Fisheries (J009997), entitled “Development of a Forest Resource Utilization System in Difficult-to-Return Zones Using Drones Equipped with 3D Scanners and Deep Learning,” funded by the Ministry of Agriculture, Forestry and Fisheries of Japan (2021–2023).

## References

Huang, Y., Ou, B., Meng, K., Yang, B., Carpenter, J., Jung, J., and Fei, S., 2024: Tree Species Classification from UAV Canopy Images with Deep Learning Models. Remote Sensing, 16, 3836.

Ma, Y., Zhao, Y., Im, J., Zhao, Y., Zhen, Z., 2024: A Deep-Learning-based Tree Species Classification for Natural Secondary Forests using Unmanned Aerial Vehicle Hyperspectral images and LiDAR. Ecological Indicators, 159, 111608.

Huang, Y., Wen, X., Gao, Y., Zhang, Y., Lin, G., 2023: Tree Species Classification in UAV Remote Sensing Images Based on Super-Resolution Reconstruction and Deep Learning. Remote Sensing, 15, 2942.

Onishi, M., Watanabe, S., Nakashima, T., Ise, T., 2022: Practicality and Robustness of Tree Species Identification Using UAV RGB Image and Deep Learning in Temperate Forest in Japan. Remote Sensing, 14, 1710.

Popp, M. R., and Kalwij, J. M., 2023: Consumer-grade UAV imagery facilitates semantic segmentation of species-rich savanna tree layers. Scientific Reports, 13, 13892.

Natesan, S., Armenakis, C., Vepakomma, U., 2019: Resnet-Based Tree Species Classification Using UAV Images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-2/W13, 475-481.

Lee, E. R., Baek, W. K., Jung, H. S., 2023: Mapping Tree Species Using CNN from Bi-Seasonal High-Resolution Drone Optic and LiDAR Data. Remote Sensing, 15, 2140.

Avtar, R., Chen, X., Fu, J., Alsulamy, S., Supe, H., Pulpadan, Y. A., Louw, A. S., Nakaji, T., 2024: Tree Species Classification by Multi-Season Collected UAV Imagery in a Mixed Cool-Temperate Mountain Forest. Remote Sensing, 16, 4060.

Zhong, H., Lin, W., Liu, H., Ma, N., Liu, K., Cao, R., Wang, T., Ren, Z., 2022: Identification of tree species based on the fusion of UAV hyperspectral image and LiDAR data in a coniferous and broad-leaved mixed forest in Northeast China. Frontiers in Plant Science, 13:964769.

Saovana, N., Yabuki, N., Fukuda, T., 2021: Automated point cloud classification using an image-based instance segmentation for structure from motion. Automation in Construction, 129, 103804.

Agisoft Metashape, https://oakcorp.net/agisoft/ (2024.01.30)

Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision. Cambridge Press.

Zhong, L., Dai, Z., Fang, P., Cao, Y., Wang, L., 2024: A Review: Tree Species Classification Based on Remote Sensing Data and Classic Deep Learning-Based Methods. Forest, 15, 852.

Simonyan, K., Zisserman, A., 2015: Very Deep Convolutional Networks for Large-Scale Image Recognition. Proc. of International Conference on Learning Representations (ICLR).

Kokkinos, Y., Margaritis, K.G., 2014: Breaking ties of plurality voting in ensembles of distributed neural network classifiers using soft max accumulations. IFIP International Conference on

Artificial Intelligence Applications and Innovations, pp. 2028, Springer.

Misra, D., Crispim-Junior, C. F., Tougne, L., 2020: Patch-based CNN evaluation for bark classification. Proc. European Conference on Computer Vision.

Veras, H. F. P., Ferreira, M. P., Cunha Neto, E. M. da, Figueiredo, E. O., Dalla Corte, A. P., Sanquetta, C. R., 2022: Fusing Multi-Season UAS Images with Convolutional Neural Networks to Map Tree Species in Amazonian Forests. Ecological Informatics, 71, 101815.