

RESEARCH ARTICLE

Large-scale, image-based tree species mapping in a tropical forest using artificial perceptual learning

Chengliang Tang¹  | María Uriarte^{1,2}  | Helen Jin³  | Douglas C. Morton⁴  | Tian Zheng¹ 

¹Department of Statistics, Columbia University, New York, NY, USA

²Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, NY, USA

³Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

⁴NASA Goddard Space Flight Center, Greenbelt, MD, USA

Correspondence

Tian Zheng

Email: tian.zheng@columbia.edu

Funding information

National Science Foundation, Grant/Award Number: NSF DEB-0620910 and NSF IIS-1741191

Handling Editor: Aaron Ellison

Abstract

1. Information about the spatial distribution of species lies at the heart of many important questions in ecology. Logistical limitations and collection biases, however, limit the availability of such data at ecologically relevant scales. Remotely sensed information can alleviate some of these concerns, but presents challenges associated with accurate species identification and limited availability of field data for validation, especially in high diversity ecosystems such as tropical forests.
2. Recent advances in machine learning offer a promising and cost-efficient approach for gathering a large amount of species distribution data from aerial photographs. Here, we propose a novel machine learning framework, artificial perceptual learning (APL), to tackle the problem of weakly supervised pixel-level mapping of tree species in forests. Challenges arise from limited availability of ground labels for tree species, lack of precise segmentation of tree canopies and misalignment between visible canopies in the aerial images and stem locations associated with ground labels. The proposed APL framework addresses these challenges by constructing a workflow using state-of-the-art machine learning algorithms.
3. We develop and illustrate the proposed framework by implementing a fine-grain mapping of three species, the palm *Prestoea acuminata* and the tree species *Cecropia schreberiana* and *Manilkara bidentata*, over a 5,000-ha area of El Yunque National Forest in Puerto Rico. These large-scale maps are based on unlabelled high-resolution aerial images of unsegmented tree canopies. Misaligned ground-based labels, available for <1% of these images, serve as the only weak supervision. APL performance is evaluated using ground-based labels and high-quality human segmentation using Amazon Mechanical Turk, and compared to a basic workflow that relies solely on labelled images.
4. Receiver operating characteristic (ROC) curves and Intersection over Union (IoU) metrics demonstrate that APL substantially outperforms the basic workflow and attains human-level cognitive economy, with 50-fold time savings. For the palm and *C. schreberiana*, the APL framework has high pixelwise accuracy and IoU with reference to human segmentations. For *M. bidentata*, APL predictions are congruent with ground-based labels. Our approach shows great potential for leveraging existing

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society

data from global forest plot networks coupled with aerial imagery to map tree species at ecologically meaningful spatial scales.

KEYWORDS

automatic species identification, machine learning, semantic segmentation, species distribution, weakly supervised learning

1 | INTRODUCTION

Characterizing the spatial distribution of species is a central goal in ecology, conservation and evolutionary biology (Gaston, 2003; Sexton et al., 2009; Wisz et al., 2013). The most common tools used for this task are species distribution models (SDMs) which link presence or abundance data with the environmental factors that are expected to determine where a species can be found (Dormann et al., 2012; Elith et al., 2011; Guisan & Thuiller, 2005; Kozak et al., 2008). These tools, however, are often limited by availability of species occurrence or abundance data collected at ecologically meaningful spatial scales, particularly in tropical areas. For tree species, SDMs often rely on herbaria data and field surveys (e.g. Muscarella et al., 2014). Yet these types of data have well-known limitations. Collections of herbaria data are driven by taxonomic questions and constrained by logistical concerns while field surveys are necessarily limited in extent and often located in areas near human habitation, potentially introducing biases in results (McMichael et al., 2017).

Remote sensing techniques such as satellite images and stereo photos provide large-scale, high-resolution data of tree canopies that allow researchers to gather information on tree species distributions across large areas (Brodrick et al., 2019; Wäldchen et al., 2018). However, processing these data presents some challenges. In principle, manual image interpretation by human experts can recognize and extract critical image features including tone or colour, shape, size, pattern, texture, shadows and context (Lillesand et al., 2015; Paine & Kiser, 2012). Manual processing, however, requires expert knowledge and can be subjective and prohibitively costly and time-consuming. Recent developments in machine learning methods offer a promising and cost-efficient approach for gathering a large amount of species occurrence and abundance data from aerial photographs by extracting the same image features used in manual interpretation. For example, edge detection calculates the colour change of neighbouring pixels and is one useful technique for identifying features such as roads (Rowe et al., 2001) and clouds (Huang et al., 2018). In combination with samples of georeferenced ground observations (i.e. training data), these methods have the potential to generate tree species spatial distribution data at large, ecologically meaningful scales.

Currently, a large number of tree species mapping studies are conducted at a spatial resolution of 10–30 m (Grabska et al., 2019; Hansen et al., 2013). These studies have successfully revealed important environmental phenomena, but are still too spatially coarse for more precise analyses of tree species distributions and related environmental processes (Wagner et al., 2019). Semantic segmentation is one way to produce detailed tree species maps using

aerial images because it enables the delineation of nonrectangular objects, such as tree canopies, in full detail (Brodrick et al., 2019). Fully Convolutional Networks (FCNs; Long et al., 2015) use classification networks as convolutions with kernels that cover the entire input regions, leading to efficient computation that is able to handle large, remote sensing datasets (Brodrick et al., 2019). In the past few years, FCN model variants such as U-Net (Ronneberger et al., 2015; Wagner et al., 2019) and Dilated Convolutions (Yu & Koltun, 2015) have greatly improved performance. Despite their excellent performance, such methods require strong supervision in the form of precisely segmented tree canopies from experts, which are costly and especially challenging to obtain for most rainforest tree species. When lacking high-quality segmented labels, available data from mapped, ground inventories of tree species offer a source of weak supervision for such a task.

Learning with weak supervision in machine learning has received increasing attention due to its broad applicability to real-world data (Zhou, 2018). Supervision from available 'ground truth' is weak when the labels are incomplete, inexact and/or inaccurate. Incomplete supervision refers to a situation where labels are only available for a small subset of training data (Zhu & Goldberg, 2009). Inexact supervision arises when, rather than pixelwise segmentation, only coarse-grained image-level labels are available (Foulds & Frank, 2010). Inaccurate supervision concerns the situation where labels constitute a noisy version of the ground truth (Ghosh et al., 2017; Karimi et al., 2020). Advances in weakly supervised learning (Goeau et al., 2017; Oliver et al., 2018; Patrini et al., 2017; Thulasidasan et al., 2019) mostly concern individual issues of weak supervision and rely on simplifying assumptions of label noises.

In this paper, we propose to tackle the task of using stem locations of tree species from a 16-ha mapped tree census as weak supervision to train a workflow for tree species mapping over a 5,000-ha area. The challenges of this task correspond to all three issues of weak supervision. First, field surveys and herbaria typically cover only a small subset of an entire species' range and tree canopies. Second, ground labels come as stem locations of different tree species, which are points in space, as opposed to precisely segmented areas of tree canopies. Third, in tropical rainforests, there is substantial misalignment between the visible canopies in the aerial images and the stem locations associated with ground labels since tree trunks often bend at sharp angles to gain greater sunlight exposure.

To address the above challenges, we developed a machine learning visual perception workflow, *artificial perceptual learning* (APL), and applied it to obtain pixelwise density predictions for three arboreal species, the Sierra palm *Prestoea acuminata* and the tree species

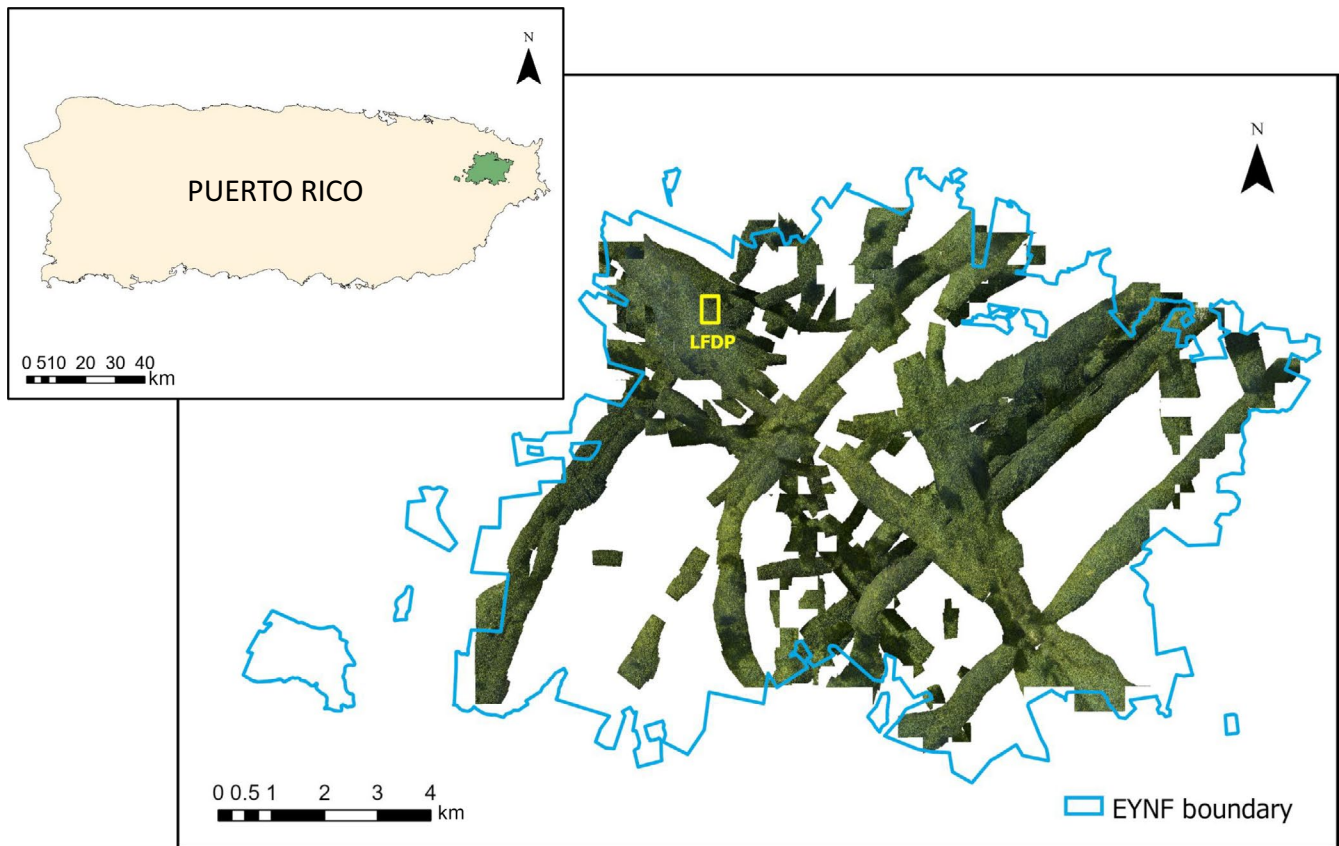


FIGURE 1 Upper left: Location of El Yunque National Forest (green area) in Puerto Rico; Lower right: G-LiHT flight lines (green area), location of the Luquillo Forest Dynamics Plot (LFDP; yellow box)

Cecropia schreberiana and *Manilkara bidentata*, in El Yunque National Forest (EYNF) in Puerto Rico (Figure 1). The proposed workflow leverages an extensive airborne collection of fine-resolution imagery over the forest. The computer vision task is to map species canopies using these unlabelled high-resolution aerial images. Imprecise (and inaccurate) ground locations of tree stems in a permanent 16-ha mapped plot (<1% of the images) serve as the only weak supervision. We compare the results of our analyses to predictions obtained using a conventional supervised learning paradigm, where noisy ground observations were simply treated as image labels. We validate the results using a set of images segmented by Amazon Mechanical Turk human annotators and cross-validation with ground labels.

2 | MATERIALS AND METHODS

2.1 | Data: Aerial images and ground labels

In March 2017, high-resolution (3 cm × 3 cm) aerial images were captured at the landscape scale (area > 5,000 ha) over EYNF in Puerto Rico by NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager (Figure 1; Cook et al., 2013). In total, this dataset consists of 962 unlabelled full-colour TIFF images, each of size 10,000 px × 10,000 px (300 m × 300 m). All images were pre-processed prior to analyses using a common shadow detection and

removal procedure (Bradski, 2008). We did not use any other data from the G-LiHT mission.

Ground labels of tree species are available for a 16-ha area, the Luquillo Forest Dynamics Plot (LFDP), covering <1% of the area captured by the aerial images (Figure 1). All stems ≥1 cm in diameter in the plot were censused in 2016, including information on species taxonomy and stem diameters, and all trees are fully georeferenced (see Thompson et al., 2002 for details). In this paper, we used ground label information from all live stems ≥20 cm diameter in the 2016 census. This size class is expected to include most individuals with visible canopies in the aerial photographs and provides more reliable supervision information for the proposed method. We also ran experiments using 15 cm and 25 cm as the threshold, and results are similar (Appendix C).

2.2 | The artificial perceptual learning workflow

Motivated by the human cognitive development process known as infant categorization, the APL framework is built with state-of-the-art machine learning algorithms as building blocks to achieve better learning efficiency. The main idea is to create a large high-quality training set of labelled image patches (e.g. 100 px × 100 px = 3 m × 3 m) by applying the APL workflow to image analyses where only weak supervision is available (e.g. noisy labels on a small subset of the

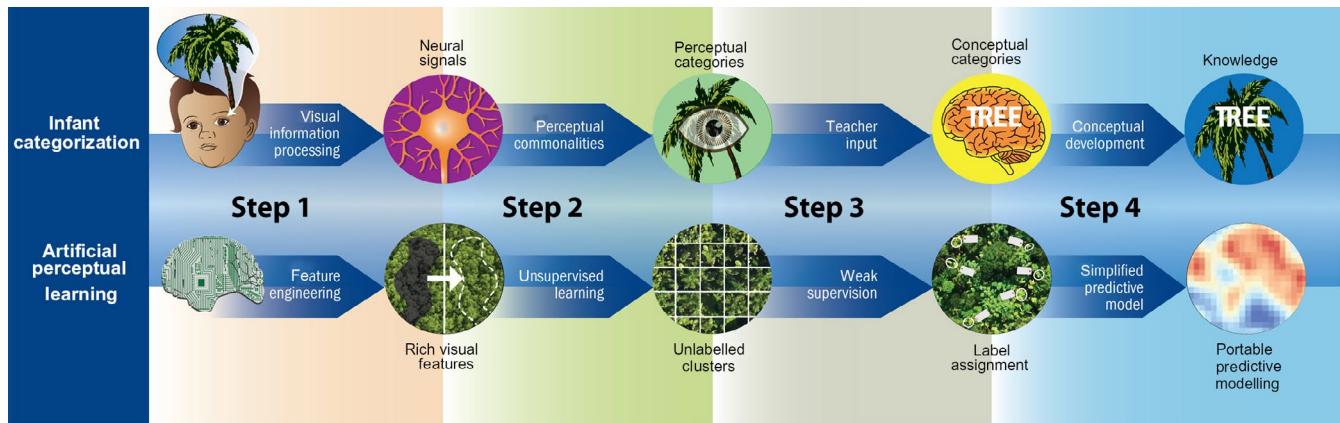


FIGURE 2 Artificial perceptual learning (APL) compared to the cognitive process of infant categorization

patches). Stable and portable predictive models can then be constructed based on this large training set using conventional supervised learning algorithms. Artificial perceptual learning is composed of the following four steps (Figure 2): (a) visual feature engineering, (b) unsupervised learning, (c) prototype-label assignment with weak supervision and (d) pixelwise classification.

Visual feature engineering, step 1 in the APL framework, is to derive what a computer ‘sees’ in an image. Here, the goal is twofold: to remove random and non-random noise, which is application specific, and to process unstructured arrays of pixels of an image into structured vectors of rich and meaningful visual features. This visual feature engineering step is similar to the bottom-up processing structure of human vision (Gibson, 1966). Visual stimuli received by our eyes are processed into neural signals by the visual nervous system and interpreted by our brains (Figure 2). In APL, existing general-purpose pre-trained deep networks in computer vision, such as AlexNet (Krizhevsky et al., 2012), VGG network (Simonyan & Zisserman, 2014) and ResNet (He et al., 2016) are well-suited for learning expressive image representations, as they have achieved excellent generalization performance in transfer learning (Ng et al., 2015; Shin et al., 2016). In particular, they can map different visual patterns, for example, colour, shape, textures, into linearly separable classes in feature space (Guérin et al., 2017; Xie et al., 2016).

Steps 2 and 3, unsupervised learning and prototype labelling with weak supervision, are the key components of the APL workflow, an idea that originates from our understanding of infant categorization (Mareschal & Quinn, 2001). Studies have shown that even infants under 1 year are able to spontaneously recognize ‘basic-level’ categories, known as perceptual categories, with little supervision (Murphy, 2004). These perceptual categories are simple groupings of visually similar stimuli. An infant assembles these grouping according to perceptual commonalities (e.g. shape, colour, etc.) without understanding their significance. These primitive perceptual categories provide the basis for a cognitive structure to incorporate supervision and form advanced and abstract categories.

In the proposed APL framework, the cognitive process of forming perceptual categories is replaced by step 2, applying unsupervised

learning algorithms to the extracted image features. Deep features from pre-trained models are expected to map visually similar image samples into proximate neighbourhoods in the feature space. Conventional clustering algorithms, such as K-means clustering, can efficiently and reasonably organize the image set into (disjoint) groups that have closely similar visual patterns.

Step 3 of the APL workflow corresponds to a parent/teacher's input in the process of infant categorization by providing imprecise information on a collection of mixed perceptual categories. With this weak supervision, infants are able to turn perceptual categories into knowledge of conceptual prototypes. Weak supervision may come in different forms depending on the setup of the task. For example, in the case of inaccurate supervision (noisy labels), a practical solution is to calculate an ‘average’ label. More specifically, for each of the unlabelled clusters (i.e. prototypes) identified in step 2, the ‘average’ label represents its relevance to the class labels of interest (e.g. target species or non-target species), quantified using weak supervision provided by the small subset of image patches with noisy labels.

The fourth and final step of the APL workflow corresponds to the post-infant conceptual development. When learning to categorize new objects, a more developed brain does not merely rely on simple perceptual categories based on characteristics. Instead, it will take advantage of previously learned concepts and category structures as meta features in a more supervised setting. Applying this idea to APL, labelled clusters (i.e. prototypes) serve as the learned knowledge. In practice, they can now be used as a training set for either conventional predictive modelling algorithms or the multi-instance learning framework.

3 | RESULTS

3.1 | Implementation of APL to case study

We used the APL framework to estimate the spatial distribution of three target species, the palm *Prestoea acuminata* and the tree species *Cecropia schreberiana* and *Manilkara bidentata* over El Yunque rainforest using all the aerial images captured by NASA G-LiHT

Airborne Imager (Cook et al., 2013). Here, we use the palm species to illustrate the major steps, but the procedure was identical for the other species. Following the APL workflow, we constructed pixelwise estimation (resolution = 10 px × 10 px, 30 cm × 30 cm) for the distribution of the three target species over the forest. Finally, we quantitatively evaluated the prediction performance of APL versus that of a basic workflow that relies solely on annotated labels using Amazon Mechanical Turk (MTurk) human segmentations (for *P. acuminata* and *C. schreberiana*) and ground observations (for all three species). We also used the MTurk segmentations to compare the APL framework with the U-Net algorithm (Wagner et al., 2019), a framework that requires strong supervision for image classification.

Ground labels from the LFDP area served as the only source of weak supervision in the APL framework. First, we split the LFDP area into 14,400 small patches (size = 100 px × 100 px = 3 m × 3 m), and mapped the ground observations onto their spatial locations. All the patches could be categorized into three types (Figure 3). The first type are the palm patches ($n = 102$), which correspond to the case when one or more observations of palm stems ≥ 20 cm diameter fall within the patch. The second type are the non-palm patches ($n = 3,136$) when all observations belong to one of more stems ≥ 20 cm diameter of a

non-palm species. The last type are the unassigned patches ($n = 11,162$), which lack any actual observation (i.e. stem ≥ 20 cm DBH) in the patch and thus, yield no information about its category. Note that this third case can occur even when there is visible foliage in the canopy, simply because we only have point locations of stems in the ground data, not canopy areas. Since the ground observations for palm trees >20 cm dbh are rather sparse over the LFDP, the last type accounts for the largest proportion of patches (11,162/14,400 $\approx$ 80%).

The image patches are the smallest analysis units in our APL workflow. The patch size is selected to be 100 px × 100 px (3 m × 3 m, about $\frac{1}{4}$ of the crown size), throughout this paper, so that each patch can be approximately associated with a single tree canopy, while still provides important texture information.

The first step was feature engineering. We fed all the 14,400 patches, regardless of their type, through the pre-trained ResNet-50 network (He et al., 2016). The output features were harvested from the last average pooling layer. In this way, we transformed all the 100 px × 100 px (3 m × 3 m) image patches into a group of partially labelled 2048-dimensional features.

Prototypes with continuous labels were derived in two steps. First, a K-means clustering model ($K = 25$) was fitted to discover

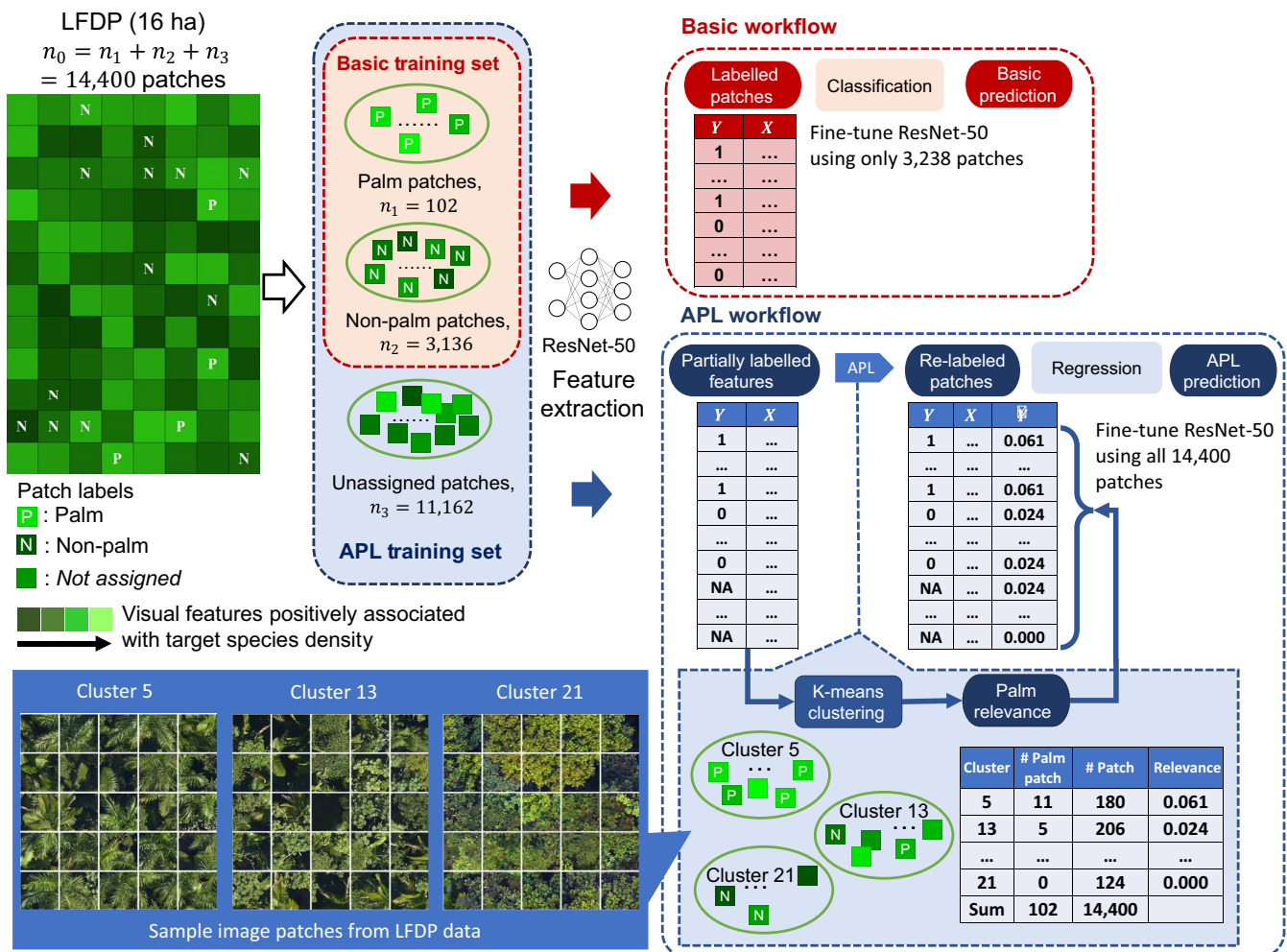


FIGURE 3 Diagram for artificial perceptual learning (APL) and basic workflow. We show the process for the sierra palm *Prestoea acuminata* for illustration purposes but these steps can be generalized to any species

prototypes composed of visually similar patches (Figure 3). This process organizes all the image patches into multiple disjoint groups according to their distances in the 2048-dimensional feature space. Once image patches had been divided into disjoint clusters, the proportion of palm patches in each cluster (palm relevance) was calculated as the number of palm patches divided by the total number of patches in each cluster. Intermediate results of prototype identification are provided in Supplementary Materials (Appendix A).

In the last step, a patch predictor was trained over visual features using a conventional supervised learning framework. To create a pixelwise prediction of species spatial distribution over the forest, we fine-tuned the ResNet-50 network (initialized with its pre-trained weights) over the relabelled patches in a multi-label regression framework, and then treated the trained network predictor as a convolutional kernel. Here, a multi-label regression was built based on the last fully connected layer with Sigmoid activation. The training hyper-parameters are 30 epochs with Adam optimizer (Kingma & Ba, 2014), batch size = 64 and learning rate = 1×10^{-4} , selected based on validation on a hold-out validation set (20% of the training data). Using a sliding window of fixed size (step size = 10 px $\times$ 10 px, window size = 100 px $\times$ 100 px), highly overlapped local patches were then fed to the fine-tuned ResNet-50 network to generate the density predictions of the three target species. Finally, we averaged the predictions on overlapped areas into a consolidated distribution prediction (resolution = 10 px $\times$ 10 px). Prediction results were stable under different choices of diameter threshold (≥ 15 , 20 and 25 cm) and number of clusters ($K = 20$, 25 and 30; Appendix C). Over the entire forest, we applied the fine-tuned network predictor to create a 10 px $\times$ 10 px species distribution map (Figure 4, Appendix B).

Artificial perceptual learning was coded in Python language and Tensorflow backend. Model training took ~1 hr using Graphics Processing Unit (GPU) on an NVIDIA Tesla V100 module with a 16 GB dedicated memory. Prediction using GPU of a single TIFF image (size: 10,000 px $\times$ 10,000 px) took approximately ~5 min.

3.2 | The basic workflow

For comparison, we constructed a basic workflow by fine-tuning the ResNet-50 network in a multi-label classification framework (using 50 epochs with the Adam optimizer, batch size = 128 and learning rate = 1×10^{-3}). This is similar to the last step in APL, but only using the raw labelled patches as training data. This workflow is rather simple and straightforward. Since there is no labelling information about the unassigned patches (Figure 3), the available training set for the basic workflow is limited to the first two patch types, which account for a small proportion of LFDP (~22%). Furthermore, the raw patch labels are of poor quality due to the mismatch between canopies (patch label) and stems (ground label).

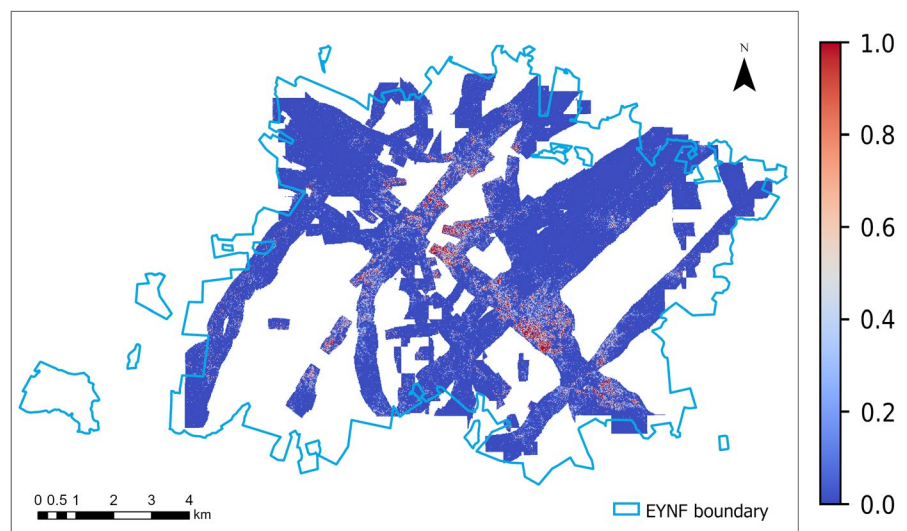
3.3 | Performance evaluation

3.3.1 | MTurk annotators

To quantitatively evaluate the performance of the APL framework, we randomly chose three sample images (size = 10,000 px $\times$ 10,000 px = 300 m $\times$ 300 m) from different locations as the test data, and then collected human segmentations from Amazon MTurk. Online MTurk workers were required to segment target tree canopies by drawing closed polygons. The polygons were then transformed into pixelwise classification of the target species. With 10 submissions collected for each image, we took their pixelwise majority vote to construct reference human classification (Appendix D). MTurk segmentations were obtained for the palm and *C. schreberiana*. For *M. bidentata*, with its thick, small simple leaves, even the most experienced forest ecologists cannot distinguish the species in 3 cm $\times$ 3 cm aerial images. This limitation precludes evaluation using MTurk annotators for *M. bidentata*.

Estimated species distributions were constructed for both the basic and APL workflow. Predictors of both these two workflows were built by fine-tuning the ResNet-50 network. Treating the calculated human

FIGURE 4 Visualization of Artificial perceptual learning (APL) predictions of the palm *Prestoea acuminata* distribution over El Yunque National Forest. The colour bar represents the estimated proportion of area covered by the target species (see Appendix B for more detail)



segmentation reference as ground truth, three performance metrics were computed. First, receiver operating characteristic curves (ROC) for the two species were visualized and the area under the ROC curve (AUC) was calculated. The ROC curve displays how the true positive rate (TPR) and the false positive rate (FPR) change with the value of the classification threshold. AUC provides an aggregate performance measure across all classification thresholds. Secondly, precision–recall curves were visualized and F1 scores were calculated. We also calculated the Intersection over Union (IoU) score, a common measure of performance in image segmentation tasks. IoU is calculated as True positives/(True positives + False positives + False negatives). Based on IoU, we chose an optimal cut-off value to transform continuous density predictions into a binary output (Appendix E).

ROC curves show that APL outperforms the basic workflow and is close in performance to that of human annotators (Figure 5). AUC for the palm with APL was 0.93 versus 0.82 with the basic framework. For *C. schreberiana*, the corresponding numbers were AUC = 0.88 for APL versus 0.67 for basic (Figure 5). Considering the skewed proportion of target and non-target species, precision–recall curves were also calculated (Figure 6). Using the optimal cut-off for maximizing IoU (Appendix E), the APL framework has pixelwise accuracy of 91.8% and IoU of 0.58. For *C. schreberiana*, the corresponding performance values are 88.0% and 0.39. F1 scores were 0.73 for the palm and 0.57 for *C. schreberiana*. Performance evaluations in terms of accuracy and IoU also reveal the significant improvement in APL over the basic workflow (Figure 7). *C. schreberiana* prediction has

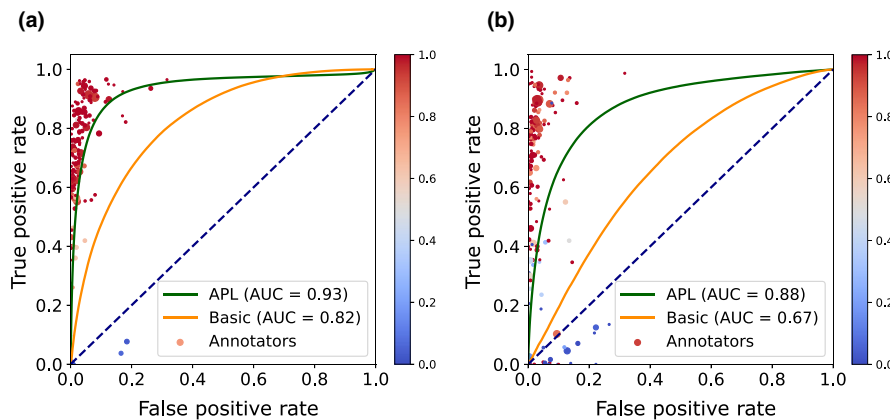


FIGURE 5 Receiver operating characteristic (ROC) curve of artificial perceptual learning (APL) and basic workflow predictions and annotator performance. (a) *Prestoea acuminata*; (b) *Cecropia schreberiana*. Point sizes represent the number of assignments submitted by each annotator. Point colours represent the acceptance rate (0–1) of each annotator (bar on the right)

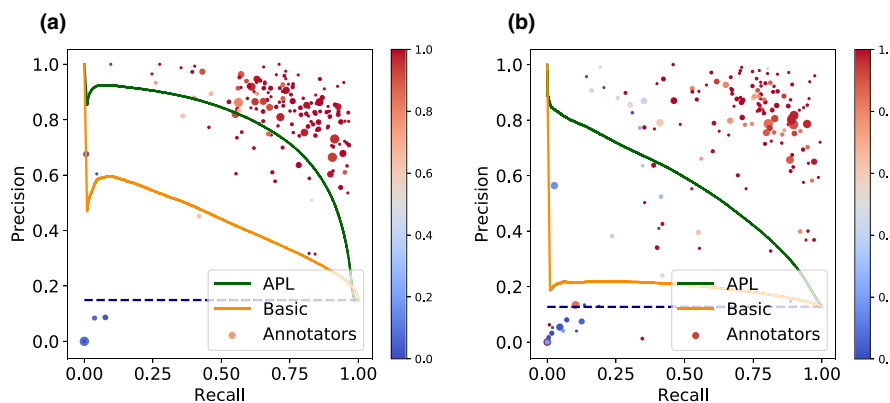


FIGURE 6 Precision–recall curve of artificial perceptual learning (APL) and basic workflow predictions and annotator performance. (a) *Prestoea acuminata*; (b) *Cecropia schreberiana*. Point sizes represent the number of assignments submitted by each annotator. Point colours represent the acceptance rate (0–1) for each annotator

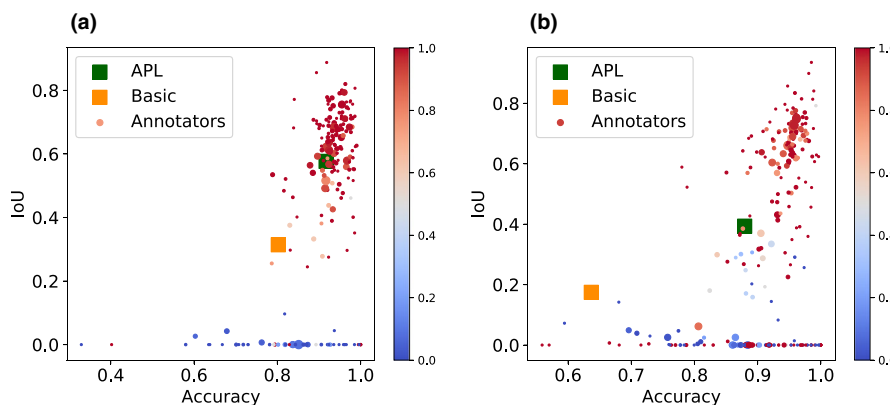
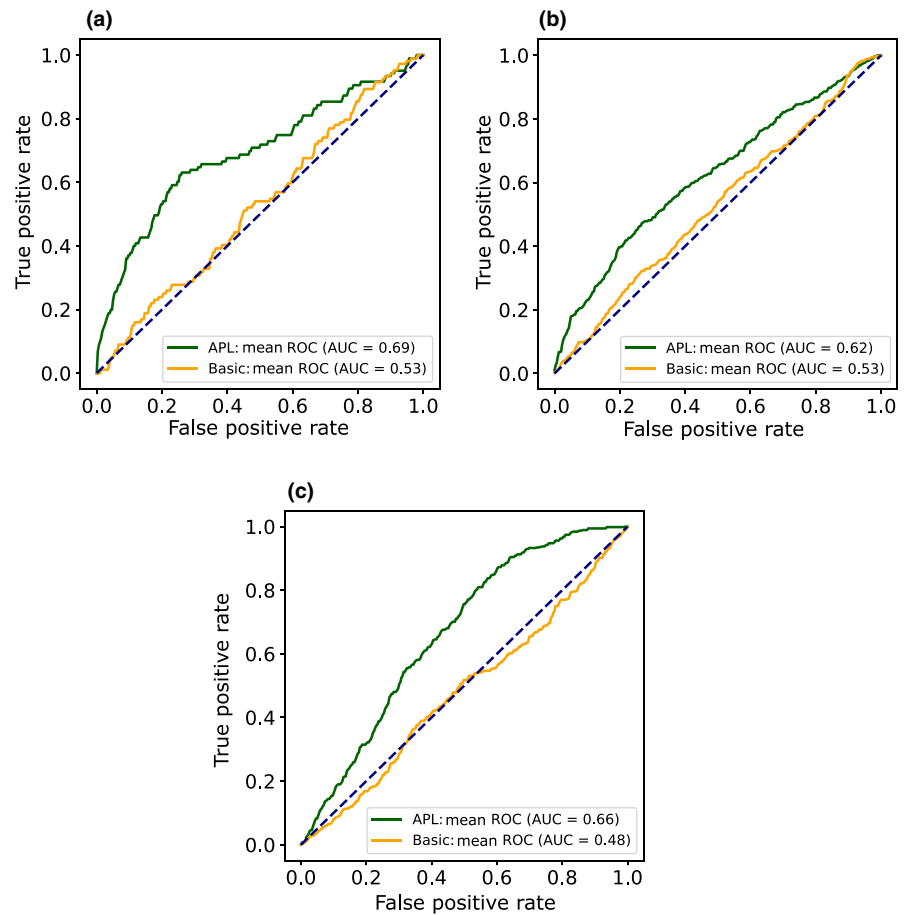


FIGURE 7 Performance evaluation for artificial perceptual learning (APL) and basic workflow: Intersection over Union (IoU) versus accuracy. (a) *Prestoea acuminata* prediction; (b) *Cecropia schreberiana* prediction. Scatter points depict annotator performance with point size proportional to the number of assignments submitted by each annotator. Point colours represent the acceptance rate of each annotator

FIGURE 8 Receiver operating characteristic (ROC) curve of artificial perceptual learning (APL) prediction over Luquillo Forest Dynamics Plot (LFDP) area by cross-validation. (a) *Prestoea acuminata*; (b) *Cecropia schreberiana*; (c) *Manilkara bidentata*



a lower performance score compared with *P. acuminata*. There are two reasons for this difference. First, *C. schreberiana* tree canopies are sparser and have long thin, white branches and they grow opportunistically in canopy openings. Therefore, their shapes are more irregular than *P. acuminata*, making weakly supervised learning more difficult. Secondly, the irregular canopy shapes also make *C. schreberiana* segmentation more difficult for MTurk annotators. As a result, the collected human reference is less reliable. We also compared the APL framework with the U-Net algorithm (Wagner et al., 2019) using strong supervision from the MTurk annotations (Appendix F). Results demonstrate that performance degradation between U-Net and our approach is not significant, showing that the APL framework efficiently extracts useful information from weak supervision.

Artificial perceptual learning also achieved substantial time savings relative to human annotators. Using a 2.4 GHz Intel Xeon processor with an NVIDIA Tesla V100 module, on average an MTurk worker took 170 s to segment a 1,000 px × 1,000 px image while the APL classifier completed this task in 3 s, a 50-fold improvement.

3.4 | Ground-based labels

To further assess the performance of APL, we evaluated predictions using ground labels. Besides providing an additional performance metric, this approach allowed us to include *M. bidentata*.

Artificial perceptual learning evaluation using ground labels was conducted using cross-validation over the LFDP area. To avoid spatial correlation between the training set and the validation set, we evenly cut the whole rectangular 16-ha area into four equal blocks and used each block as the unit fold in the cross-validation (Appendix E). Specifically, in each round, we trained the predictor over three blocks, and then evaluated the trained predictor over the remaining block. After four rounds, we combined the output density predictions on the four test folds and compared them to ground observations using ROC and AUC (Appendix E).

Over the LFDP area, the test AUCs for the three species are 0.69 (*P. acuminata*), 0.62 (*C. schreberiana*) and 0.66 (*M. bidentata*; Figure 8). The basic workflow is rather close to random guessing with AUCs around 0.5, due to the high noise level in the data and the poor generalization of the model.

4 | DISCUSSION

The explosion in the availability of remotely sensed data (e.g. aerial images, Light Detection and Ranging) offer unprecedented opportunities to map tree species across large areas and at ecologically meaningful spatial scales (Brodrick et al., 2019). Despite the increasing application of machine learning algorithms to image-based species identification over the last few years (see review in

Fassnacht et al., 2016; Wäldchen et al., 2018), tree species mapping using semantic segmentation of canopies from remotely sensed images still presents a number of challenges. One of the most critical challenges is the limited availability of training data. For this reason, tree species mapping using machine learning has relied on human manual segmentation to generate training datasets (Li et al., 2019; Wagner et al., 2019). Manual processing, however, requires expert knowledge and can be subjective, prohibitively costly and time-consuming. An alternative is to use data from mapped forest inventories but typically, these data cover only a small subset of an entire species' range. Furthermore, tree canopies, the segmented objects of interest, do not align with pixels from aerial images and significant spatial misalignment can occur between tree canopies and ground stem observations, further complicating the use of mapped plot data. These limitations are particularly marked in tropical rainforests, where complex, multilayered canopies and high species diversity further complicate species identification from remotely sensed data.

The proposed APL workflow takes advantage of recent developments in the field of machine learning under weak supervision to address these limitations. Rather than relying on manual segmentations of tree canopies or annotated labels for training, our workflow rests on the idea that images that contain similar densities of a given target species will map into proximate neighbourhoods in the feature space. This step mimics the way an infant assembles objects into grouping according to perceptual commonalities (e.g. shape, colour, etc.) even in the absence of a label for supervision (i.e. name of object; Mareschal & Quinn, 2001; Murphy, 2004). Evaluation using human annotators show that APL attains human cognitive ability and at the same time, outperforms a basic workflow that relies solely on labelled images (i.e. images with known species) for training. The algorithm can also capture species distributions derived from ground data and its performance is similar to algorithms that require strong supervision from human segmentations.

Climate models predict that some of the fastest increases in temperature are expected to occur in tropical regions (Anderson, 2011; Christensen et al., 2007; Diffenbaugh & Scherer, 2011; Mora et al., 2013) and precipitation patterns are also expected to shift, towards drier conditions in many tropical regions (Dai, 2013; Feng et al., 2013; Mora et al., 2013; Neelin et al., 2006). These shifts in climate are likely to influence the distribution of tree species. Improving our understanding of the factors that drive the spatial distribution of tree is particularly critical for tropical regions, where national forest inventories are rare and herbaria collections are limited. Yet, most tree species classification studies have been conducted in North America and Europe (Fassnacht et al., 2016). Remotely sensed data coupled with extensive forest plot networks have the potential to advance our understanding of the factors that influence the spatial distribution of tree species areas across large areas. The APL workflow offers a way to leverage data from extensive, global mapped forest plot networks (e.g. Smithsonian ForestGeo <https://forestgeo.si.edu/what-forestgeo>) with high-resolution aerial images (e.g. Worldview, Quickbird, high-resolution drone imagery) or other remotely sensed data (e.g.

LiDAR, hyperspectral images) to achieve this task in a computationally efficient way.

The proposed APL framework provides a general workflow to address studies where limited data necessitate weak supervision, a situation that is common in ecology. In this paper, we use the workflow to construct a scalable pixelwise segmentation of species distribution with limited ground observations. However, there still exist a few limitations in the our approach. The first limitation is the lack of continuity in the prediction output. In the implementation of APL, the patch predictor was applied as a convolutional kernel to generate species distributions. As a result, the predicted target species distribution would highly depend on local visual features, and fall short of ensuring spatial continuity at a larger scale. Secondly, multiple species coexist and their canopies mix significantly. The current workflow predicts the distribution of three tree species in a multi-label regression framework, which might ignore the local interactions of different species. The last limitation lies in prediction inaccuracies around canopy boundaries, which are evident in the majority of APL segmentation. This limitation is rather specific to our task settings, and is unavoidable due to the lack of boundary information in the training set.

There are a number of extensions that could address the above limitations. The first is to incorporate spatial priors into the predictive model and generate spatially continuous predictions. Another potential extension is to build the APL workflow over multiple data sources. For example, a growing research area is to fuse LiDAR point clouds with aerial images for species identification (Fassnacht et al., 2016). The output data fusion can add a depth dimension to the 2D images, which will help the segmentation of canopy boundaries in our task. Finally, the identifiability of species canopies is mostly controlled by image resolution. Future advances in remote sensing technologies will improve both data availability and quality, facilitating species identification with the proposed APL workflow.

5 | CONCLUSIONS

In this paper, we tackle the problem of weakly supervised tree species mapping from aerial photographs with artificial perceptual learning. Within this framework, state-of-the-art machine learning algorithms are employed as building blocks to mimic the early stage of human concept development. To validate the proposed framework, experiments were conducted over a wide-field unlabelled ecological remote sensing data, and MTurk human segmentations were collected for performance evaluation. Results show our framework can attain human-level cognitive economy with substantial time savings. Further work is needed to incorporate spatial continuity and species interdependence into the application of the workflow in ecological studies.

ACKNOWLEDGEMENTS

Collection of tree data for the LFDp was supported by award NSF DEB-0620910 to the Luquillo LTER. The International Institute of

Tropical Forestry (USDA Forest Service) provided logistical support for the G-LiHT flights.

AUTHORS' CONTRIBUTIONS

C.T., T.Z. and M.U. conceived the idea for the paper and wrote the manuscript; C.T., H.J. and T.Z. developed and implemented the APL framework; M.U. collected ground label data and D.C.M. collected aerial images. All the authors contributed critically to the draft and gave final approval for publication.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.13549>.

DATA AVAILABILITY STATEMENT

Aerial images are available at <https://gliht.gsfc.nasa.gov/index.php?section=34>. Ground label data for the Luquillo Forest Dynamics Plot are available at <https://luq.lter.network/data/luqmetadat> a119. APL codes and sample data for tree segmentation are available at GitHub and can be accessed through Zenodo <http://doi.org/10.5281/zenodo.4391471> (Tang et al., 2020).

ORCID

Chengliang Tang  <https://orcid.org/0000-0002-3691-0263>

María Uriarte  <https://orcid.org/0000-0002-0484-0758>

Helen Jin  <https://orcid.org/0000-0002-8654-7909>

Douglas C. Morton  <https://orcid.org/0000-0003-2226-1124>

Tian Zheng  <https://orcid.org/0000-0003-4889-0391>

REFERENCES

- Anderson, B. T. (2011). Near-term increase in frequency of seasonal temperature extremes prior to the 2°C global warming target. *Climatic Change*, 108(3), 581–589. <https://doi.org/10.1007/s10584-011-0196-4>
- Bradski, G. A. K. (2008). *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media Inc.
- Brodrick, P. G., Davies, A. B., & Asner, G. P. (2019). Uncovering ecological patterns with convolutional neural networks. *Trends in Ecology & Evolution*, 34(8), 734–745. <https://doi.org/10.1016/j.tree.2019.03.006>
- Christensen, J. H., Hewitson, B., Busuioac, A., Chen, A., Gao, X., Held, R., Jones, R., Kolli, R. K., Kwon, W. K., Laprise, R., & Magaña Rueda, V. (2007). Regional climate projections. In *Climate Change, 2007: The Physical Science Basis. Contribution of Working group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Chapter 11* (pp. 847–940). University Press.
- Cook, B., Corp, L., Nelson, R., Middleton, E., Morton, D., McCorkel, J., Masek, J., Ranson, K., Ly, V., & Montesano, P. (2013). NASA Goddard's LiDAR, hyperspectral and thermal (G-LiHT) airborne imager. *Remote Sensing*, 5(8), 4045–4066. <https://doi.org/10.3390/rs5084045>
- Dai, A. (2013). Increasing drought under global warming in observations and models. *Nature Climate Change*. <https://doi.org/10.1038/nclimate1633>
- Diffenbaugh, N. S., & Scherer, M. (2011). Observational and model evidence of global emergence of permanent, unprecedented heat in the 20(th) and 21(st) centuries. *Climatic Change*, 107(3–4), 615–624. <https://doi.org/10.1007/s10584-011-0112-y>
- Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., & Singer, A. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography*, 39(12), 2119–2131. <https://doi.org/10.1111/j.1365-2699.2011.02659.x>
- Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1), 43–57. <https://doi.org/10.1111/j.1472-4642.2010.00725.x>
- Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., Straub, C., & Ghosh, A. (2016). Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*, 186, 64–87. <https://doi.org/10.1016/j.rse.2016.08.013>
- Feng, X., Porporato, A., & Rodriguez-Iturbe, I. (2013). Changes in rainfall seasonality in the tropics. *Nature Climate Change*, 3, 811–815. <https://doi.org/10.1038/nclimate1907>
- Foulds, J. R., & Frank, E. (2010). A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1), 1–25. <https://doi.org/10.1017/S026988890999035X>
- Gaston, K. J. (2003). *The structure and dynamics of geographic ranges*. Oxford University Press.
- Ghosh, A., Kumar, H., & Sastry, P. (2017). Robust loss functions under label noise for deep neural networks. *arXiv preprint arXiv:1712.09482*.
- Gibson, J. J. (1966). The senses considered as perceptual systems. Houghton Mifflin.
- Goeau, H., Bonnet, P., & Joly, A. (2017). Plant identification based on noisy web data: The amazing performance of deep learning (LifeCLEF 2017).
- Grabska, E., Hostert, P., Pflugmacher, D., & Ostapowicz, K. (2019). Forest stand species mapping using the Sentinel-2 time series. *Remote Sensing*, 11(10), 1197. <https://doi.org/10.3390/rs11101197>
- Guérin, J., Gibaru, O., Thiery, S., & Nyiri, E. (2017). CNN features are also great at unsupervised classification. *arXiv preprint arXiv:1707.01700*.
- Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8(9), 993–1009. <https://doi.org/10.1111/j.1461-0248.2005.00792.x>
- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Loveland, T. R. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160), 850–853.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Huang, W., Wang, Y., & Chen, X. (2018). Cloud detection for high-resolution remote-sensing images of urban areas using colour and edge features based on dual-colour models. *International Journal of Remote Sensing*, 39(20), 6657–6675. <https://doi.org/10.1080/01431161.2018.1466069>
- Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65, 101759. <https://doi.org/10.1016/j.media.2020.101759>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kozak, K. H., Graham, C. H., & Wiens, J. J. (2008). Integrating GIS-based environmental data into evolutionary biology. *Trends in Ecology & Evolution*, 23(3), 141–148. <https://doi.org/10.1016/j.tree.2008.02.001>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks*. Paper presented at the Advances in neural information processing systems.
- Li, W. J., Dong, R. M., Fu, H. H., & Yu, L. (2019). Large-scale oil palm tree detection from high-resolution satellite images using two-stage

- convolutional neural networks. *Remote Sensing*, 11(1). <https://doi.org/10.3390/rs11010011>
- Lillesand, T., Kiefer, R. W., & Chipman, J. (2015). *Remote sensing and image interpretation*. John Wiley & Sons.
- Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully convolutional networks for semantic segmentation*. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.
- Mareschal, D., Quinn, P. C. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, 5(10), 443–450. [https://doi.org/10.1016/S1364-6613\(00\)01752-6](https://doi.org/10.1016/S1364-6613(00)01752-6)
- McMichael, C. N. H., Matthews-Bird, F., Farfan-Rios, W., & Feeley, K. J. (2017). Ancient human disturbances may be skewing our understanding of Amazonian forests. *Proceedings of the National Academy of Sciences of the United States of America*, 114(3), 522–527. <https://doi.org/10.1073/pnas.1614577114>
- Mora, C., Frazier, A. G., Longman, R. J., Dacks, R. S., Walton, M. M., Tong, E. J., Sanchez, J. J., Kaiser, L. R., Stender, Y. O., Anderson, J. M., Ambrosino, C. M., Fernandez-Silva, I., Giuseffi, L. M., & Giambelluca, T. W. (2013). The projected timing of climate departure from recent variability. *Nature*, 502(7470), 183–187. <https://doi.org/10.1038/nature12540>
- Murphy, G. (2004). *The big book of concepts*. MIT press.
- Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxentecological niche models. *Methods in Ecology and Evolution*, 5(11), 1198–1205. <https://doi.org/10.1111/2041-210X.12261>
- Neelin, J. D., Munnich, M., Su, H., Meyerson, J. E., & Holloway, C. E. (2006). Tropical drying trends in global warming models and observations. *Proceedings of the National Academy of Sciences of the United States of America*, 103(16), 6110–6115. <https://doi.org/10.1073/pnas.0601798103>
- Ng, H.-W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). *Deep learning for emotion recognition on small datasets using transfer learning*. Paper presented at the Proceedings of the 2015 ACM on international conference on multimodal interaction.
- Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., & Goodfellow, I. (2018). *Realistic evaluation of deep semi-supervised learning algorithms*. Paper presented at the Advances in neural information processing systems.
- Paine, D. P., & Kiser, J. D. (2012). *Aerial photography and image interpretation*. John Wiley & Sons.
- Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., & Qu, L. (2017). *Making deep neural networks robust to label noise: A loss correction approach*. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*. Paper presented at the International Conference on Medical image computing and computer-assisted intervention.
- Rowe, N. C., Grewe, L. I. (2001). Change detection for linear features in aerial photographs using edge-finding. *IEEE Transactions on Geoscience and Remote Sensing*, 39(7), 1608–1612. <https://doi.org/10.1109/36.934092>
- Sexton, J. P., McIntyre, P. J., Angert, A. L., & Rice, K. J. (2009). Evolution and ecology of species range limits. *Annual Review of Ecology and Systematics*, 40, 415–436. <https://doi.org/10.1146/annurev.ecolsys.110308.120317>
- Shin, H.-C., Roth, H. R., Gao, M., Lu, L. E., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285–1298. <https://doi.org/10.1109/TMI.2016.2528162>
- Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint: arXiv:1409.1556*.
- Tang, C., Uriarte, M., Jin, H., Morton, D., & Zheng, T. (2020). Data from: Large-scale, image-based tree species mapping in a tropical forest using artificial perceptual learning. *Zenodo*, <https://doi.org/10.5281/zenodo.4391471>
- Thompson, J., Brokaw, N., Zimmerman, J. K., Waide, R. B., Everham, E. M., Lodge, D. J., Taylor, C. M., García-Montiel, D., & Fluet, M. (2002). Land use history, environment, and tree composition in a tropical forest. *Ecological Applications*, 12(5), 1344–1363. [https://doi.org/10.1890/1051-0761\(2002\)012\[1344:luheat\]2.0.co;2](https://doi.org/10.1890/1051-0761(2002)012[1344:luheat]2.0.co;2)
- Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., & Mohd-Yusof, J. (2019). Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*.
- Wagner, F. H., Sanchez, A., Tarabalka, Y., Lotte, R. G., Ferreira, M. P., Aida, M. P. M., Gloor, E., Phillips, O. L., & Aragão, L. E. O. C. (2019). Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. *Remote Sensing in Ecology and Conservation*, 5(4), 360–375. <https://doi.org/10.1002/rse2.111>
- Wäldchen, J., Mäder, P., & Cooper, N. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11), 2216–2225. <https://doi.org/10.1111/2041-210X.13075>
- Wis, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J.-A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N. M., ... Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, 88(1), 15–30. <https://doi.org/10.1111/j.1469-185X.2012.00235.x>
- Xie, J., Girshick, R., & Farhadi, A. (2016). *Unsupervised deep embedding for clustering analysis*. Paper presented at the International conference on machine learning.
- Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions.
- Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53. <https://doi.org/10.1093/nsr/nwx106>
- Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1), 1–130. <https://doi.org/10.2200/S00196ED1V01Y200906AIM006>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Tang C, Uriarte M, Jin H, Morton DC, Zheng T. Large-scale, image-based tree species mapping in a tropical forest using artificial perceptual learning. *Methods Ecol Evol*. 2021;12:608–618. <https://doi.org/10.1111/2041-210X.13549>