

RandCrowns: A Quantitative Metric for Imprecisely Labeled Tree Crown Delineation

Dylan Stewart, *Member, IEEE*, Alina Zare, *Senior Member, IEEE*, Sergio Marconi, Ben G. Weinstein, Ethan P. White, Sarah J. Graves, Stephanie A. Bohlman, Aditya Singh

Abstract—Supervised methods for object delineation in remote sensing require labeled ground-truth data. Gathering sufficient high quality ground-truth data is difficult, especially when targets are of irregular shape or difficult to distinguish from background or neighboring objects. Tree crown delineation provides key information from remote sensing images for forestry, ecology, and management. However, tree crowns in remote sensing imagery are often difficult to label and annotate due to irregular shape, overlapping canopies, shadowing, and indistinct edges. There are also multiple approaches to annotation in this field (e.g., rectangular boxes vs. convex polygons) that further contribute to annotation imprecision. However, current evaluation methods do not account for this uncertainty in annotations, and quantitative metrics for evaluation can vary across multiple annotators. In this paper, we address these limitations by developing an adaptation of the Rand index for weakly-labeled crown delineation that we call RandCrowns. Our new RandCrowns evaluation metric provides a method to appropriately evaluate delineated tree crowns while taking into account imprecision in the ground-truth delineations. The RandCrowns metric reformulates the Rand index by adjusting the areas over which

Dylan Stewart and Alina Zare are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (email: azare@ece.ufl.edu).

Ben Weinstein and Ethan White are with the Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL 32611 USA.

Sarah Graves is with the Nelson Institute for Environmental Studies, University of Wisconsin-Madison, Madison, WI 53706 USA.

Sergio Marconi and Stephanie Bohlman are with the School of Forest, Fisheries, and Geomatics Sciences, University of Florida, Gainesville, FL 32611 USA.

Aditya Singh is with the Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611 USA.

each term of the index is computed to account for uncertain and imprecise object delineation labels. Quantitative comparisons to the commonly used intersection over union method shows a decrease in the variance generated by differences among multiple annotators. Combined with qualitative examples, our results suggest that the RandCrowns metric is more robust for scoring target delineations in the presence of uncertainty and imprecision in annotations that are inherent to tree crown delineation.

Index Terms—Remote sensing, Tree crown delineation, Imprecise labels, Quantitative evaluation

I. INTRODUCTION

IDENTIFYING trees in remote sensing imagery is essential to understanding individual-level ecological processes at large scales. Delineating individual tree crowns is required for research on topics ranging from plant traits [1], to carbon sequestration [2], [3], [4], to biodiversity assessment [5], [6], [7]. As a result, an array of methods have been developed for converting a variety of remote sensing data into information on individual trees. These include LiDAR-based methods that use three-dimensional point cloud representations of forest structure to cluster points into individual trees. These LiDAR-based methods include counting trees within segments by finding the number of unique Gaussian distributions needed to recreate the LiDAR data with maximum correlation [8], splitting segments which harbor multiple trees into individual delineations by finding saddle points in 2D contour maps [9], and merging watershed segments to delineate individual trees using a minimum

spanning forest [10]. There are also methods based on RGB imagery that identify groups of pixels that match the appearance of a tree from above, including a convolutional neural network pretrained on tens of millions of LiDAR-annotated crowns and fine-tuned using hand-labelled crowns in the RGB imagery [11] and a multi-sensor fusion approach of RGB images with LiDAR intensity and hand-crafted features [12]. Given the large number of methods currently available for delineating tree crowns, a key question is how best to evaluate the performance of different approaches.

Evaluation of crown delineation algorithms is currently accomplished using a variety of methods including: 1) comparisons of how many trees are delineated in a plot to the number counted in the field [13]; 2) assessment of whether individual delineated crowns are associated with a field mapped stem [14]; and 3) direct comparison of how similar delineated crowns are to expert labeled crowns [11]. Human labeled crowns are generated either by visually identifying crowns on the remote sensing imagery or by performing this task while in the field observing the trees being labeled [15]. Direct comparisons of algorithmic and human delineated crowns provides the most challenging assessment of the underlying methods. Comparisons between labeled and delineated crowns are often made based on the Intersection over Union (IoU) score, also known as the Jaccard index [16]. The IoU is the ratio of the area encompassed by the intersection of two sets divided by the union of the pair. Another related method to assess agreement between multiple annotations is the Rand index [17]. The Rand index relates the ratio of pairs of points that are grouped into the same segment (or delineation) to the total number of labeled pairs. These fundamental measures can also be converted to recall (and precision if all crowns are delineated) using an IoU/Rand index threshold.

While IoU is a common measure of overlap or similarity when delineating objects, it assumes that the ground-truth is a precise representation of the true crown. As a result, it does not account for the potentially large measurement error in human-labeled crown boundaries. Boundaries of

tree crowns are difficult for humans to delineate due to imprecise boundaries, complex three dimensional structure and object occlusion. Especially in dense forests, it can be difficult to determine whether branches belong to a target crown or neighboring crown. When relying solely on remote sensing imagery, it can also be difficult to tell the difference between single large trees and multiple smaller trees because of how small trees may cluster together, or how large trees may have multiple clusters of branches [4]. This is likely to result in significant variation in ground-truth labels depending on the person doing the labeling and the labeling method. Some of these issues can be mitigated by labeling the imagery while observing the tree in the field [15]. However, there is likely significant variation among observers in the specific crowns delimited on the scale of a few meters. Additionally, field based labeling is time consuming, making labeling based only on imagery alone an important tool. This measurement error (or “weak labeling”) influences the IoU score and current methods do not address this challenge.

The issue with evaluating weakly labeled segmentations is a common challenge in remote sensing and image segmentation more broadly [18]. There are a variety of approaches to addressing measurement error for semantic segmentation. Global and Local Consistency Error (GCE and LCE respectively) compute changes in segmentation membership across multiple partitionings [19]. The only cases where GCE and LCE award zero error are: each segment in an image consists of a single pixel or the entire image contains one segment. Variation of Information (VOI) measures the mutual information between two segmentations [20]. VOI is most appropriate when segment membership probabilities are assigned to each pixel. The Probabilistic Rand Index alters the Rand index by scaling correct pairs by their average number of occurrences within multiple ground-truth labelings [21]. While these methods attempt to address variance across multiple partitionings of semantic labels, none of these approaches directly address scoring in relation to imprecise and weakly-labeled ground-truth.

In this paper we develop a new metric, RandCrowns (RC), for evaluating overlap between algorithmic and human delineated tree crowns that takes into account the specific kinds of uncertainty and imprecision associated with tree crown delineation. Due to the imprecision and uncertainty of the desired delineations, we refer to the desired tree crown not as ground-truth, but as a desired target. This approach builds on the Rand index by using “boundaries”, or buffers, around the desired crown that provide a cushion to accommodate variation in the labels. The goal of the RandCrowns measure is to appropriately score delineations which encompass a desired crown regardless of realistic differences in labels that do not meaningfully change the labels provided. It is important to note that a relationship exists between appropriately scoring a variety of candidate delineations and assigning large scores to those with questionably large differences. This is a balance that can be controlled in our measure by modifying parameter settings. Other current methods do not allow for any balance of this kind. Although our method does require suitable selection of parameters, it is a tradeoff for an overall improvement in reliability. We evaluate this measure and compare it to the most commonly used metric for individual crown level comparisons, the IoU, using a dataset of tree crowns annotated by multiple human labelers. We find that when properly parameterized, the RandCrowns measure results in reduced variance in score resulting from differences among labelers and therefore provides a more robust approach to assessing crown delineation algorithms at the fundamental scale of the individual tree. While we develop and apply the RandCrowns measure to tree crown delineation, it can be used to analyze performance for any segmentation application whereby labels contain uncertainty (as is the case in many remote sensing problems).

II. RANDCROWNS

The traditional Rand index (RI) is a measure of similarity between two partitionings of data. The RI compares the ratio of how many pairs of points in

two partitionings agree to how many pairs of points disagree,

$$\text{Rand} = \frac{a' + b'}{a' + b' + c' + d'}, \quad (1)$$

where a' is the number of pairs of points that are found in the same segment in both partitions and b' is the number of pairs of points that are found in different segments in both partitions. The terms c' and d' are the number of pairs of points that are together in one partitioning but not in the other and vice versa. This is similar to the IoU metric which is often used to measure the overlap between predicted and observed tree crowns. However, IoU does not consider a term analogous to b' , nor does the IoU account for any uncertainty. IoU can be written with the terms from the RI as,

$$\text{IoU} = \frac{a'}{a' + c' + d'}. \quad (2)$$

The original RI, like the IoU, is simply a comparison measure between any pair of partitionings (delineations). When evaluating crown delineations, this overlap between the predicted delineation and the ground-truth represents how accurately the delineation performed. As discussed above, ground-truth crowns are typically delineated either in the field (while looking at remote sensing imagery) or by labeling remote sensing imagery alone (without field observation). If these ground-truth crowns were delineated without error then either the RI or the IoU would provide consistent measures of the individual level accuracy of the crown delineation. However, the difficulties in identifying the edges of tree crowns result in imprecision in the ground-truth labels and significant variation between human annotators when labeling the same tree crowns. Therefore, we assume that there is uncertainty to each delineated crown.

For each tree, the region drawn by a human annotator that encompasses the area of a tree may be larger or smaller than the actual canopy. This region, known as the desired target, is defined as in Equation 3,

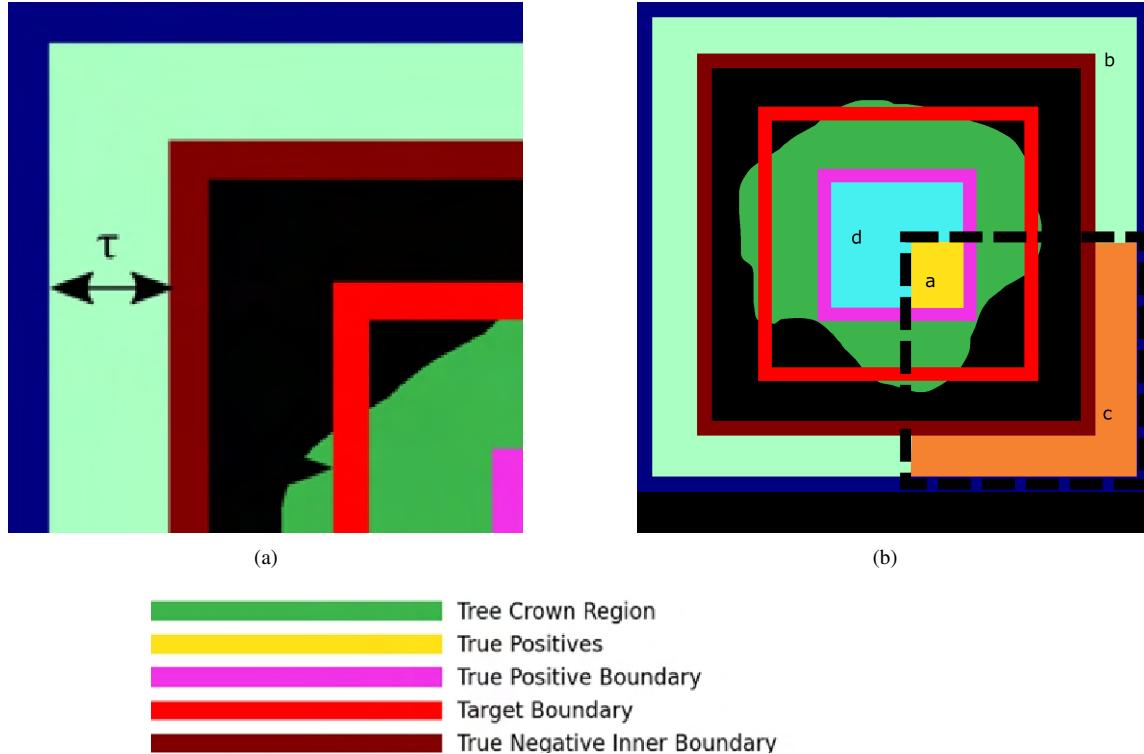


Fig. 1. A diagram of the RandCrowns regions and corresponding boundaries. An example crown is shown in green. In Figure 1a, the desired target from an annotator is shown in red and the length and height are denoted l_T and h_T respectively. The true positive boundary is within the desired target by 2α pixels in both length and height. The yellow region within the true positive boundary represents the true positive region (denoted “a” in the Rand equation) and is labeled R_a . The true negative inner boundary is shown in maroon and is larger by 2ω in both length and height than the desired target. A region that is ignored in the computation of the RandCrowns measure is set based on the true negative inner boundary, R_o . Outside of the true negative inner boundary is the true negative outer boundary shown in navy. The true negative outer boundary envelops the true negative region, R_b (denoted as “b” in the Rand equation), to be 2τ larger than the true negative inner boundary in both length and height. It is important to note within the red delineation there are some gaps where there is no crown, this represents the imprecise labeling for an imperfect “true” crown delineation. In Figure 1b, the black box is a sample delineation. The true positives are shown in the yellow highlighted region (labelled “a”); this is the intersection of the delineation and the true positive boundary. The true negatives are the mint highlighted region (labelled “b”). True negatives are the difference between the true negative outer boundary and the true negative inner boundary and the delineation. The false positives are in the orange highlighted region (labelled “c”). This region shows the portion of the delineation that is contained in R_b . The false negatives are in the cyan region (labelled “d”). This region shows the difference between the true positive boundary and the delineation.

$$T = \left\{ (x, y) \in \mathbb{R}^2 : |x - x_T| \leq \frac{l_T}{2}, |y - y_T| \leq \frac{h_T}{2} \right\}, \quad (3)$$

where (x_T, y_T) are the center coordinates of the desired target and l_T and h_T are the length and

height respectively. To deal with the imprecision of this labeling, in the RandCrowns measure, the true positive region (R_a , analogous to a') is concentric with the desired target and spans a smaller, less uncertain area within the annotated region. Due to the possibility of delineations being larger than the actual tree, the true negative region (R_b , analogous

$$D = \left\{ (x, y) \in \mathbb{R}^2 : |x - x_D| \leq \frac{l_D}{2}, |y - y_D| \leq \frac{h_D}{2} \right\} \quad (4)$$

$$R_a = \left\{ (x, y) \in \mathbb{R}^2 : |x - x_T| \leq \frac{l_T}{2} - \alpha, |y - y_T| \leq \frac{h_T}{2} - \alpha \right\} \quad (5)$$

$$R_o = \left\{ (x, y) \in \mathbb{R}^2 : \frac{l_T}{2} < |x - x_T| \leq \frac{l_T}{2} + \omega, \frac{h_T}{2} < |y - y_T| \leq \frac{h_T}{2} + \omega \right\} \quad (6)$$

$$|R_a| = (h_T + 2\omega + 2\tau)(l_T + 2\omega + 2\tau) - h_T l_T \quad (7)$$

$$R_e = \left\{ (x, y) \in \mathbb{R}^2 : \frac{l_T}{2} + \omega < |x - x_T| \leq \frac{l_T}{2} + \omega + \gamma\tau, \frac{h_T}{2} + \omega < |y - y_T| \leq \frac{h_T}{2} + \omega + \gamma\tau \right\} \quad (8)$$

to b') is concentric with the desired target but spans an area that is buffered outside of the desired boundary.

Consider a delineation, D , where the center of the delineation is indicated by (x_D, y_D) with length and height being l_D and h_D , respectively as shown in Equation 4. To quantify the relationship between the delineated and desired target, we modify each term of the Rand index to be applied to the true positive and true negative regions and denote this adaptation as the RandCrowns score. The regions for the RandCrowns index are shown in Figure 1.

For a given desired target, the true positive region, R_a , is set based on a parameter, α , as shown in Equation 5, with $\alpha > 0$. The true positive region is within the desired target and is a refined area of the tree crown for which there is more certainty that the crown is encompassed there. Outside of the desired target boundary there is an outer region.

The outer region represents the area of uncertainty that the desired target could extend to, but with high certainty does not extend outside of this area. The outer region, R_o , is computed based on the parameter, ω , and the desired target as shown in Equation 6, with $\omega > 0$. The outer region extends outside of the desired target and is used to produce the true negative region.

To properly penalize delineations that are much larger than the tree crown, the area of the true negative region is set to be a multiple, γ , of the true positive region area. Given the area of the

true positive region and the length and height of the outer region, a direct solution to produce the true negative outer boundary parameter relies on the quadratic equation in Equation 7, where $|R_a|$ denotes the area of the true positive region. The difference between the edge length and height and that of the outer region (τ in Figure 1) can be used to set the area of the true negative region, R_b . Solving for τ and adding multiples of τ to the length and height of the outer region produces an intermediate edge region in Equation 8.

In order to appropriately score large delineations that contain portions of the inner region and extend outside the true negative outer boundary, we extend the edge region for delineations that satisfy the condition: $R_e^* = R_e \cup D$. If the delineation does not extend outside of the true negative region, the region is not modified. The true negative region is computed based on the edge and outer regions: $R_b = R_e^* \setminus R_o$. Based on the solution for the edge region, changing γ in Equation 8 produces an area ratio in Equation 9,

$$\frac{|R_b|}{|R_a|} = \gamma. \quad (9)$$

For irregularly shaped desired targets, the true negative outer boundary can be set heuristically through relating the areas of the true positive and resultant true negative region. The true negative region is set by buffering the outer region iteratively and checking the area ratio of the produced true

negative region to the true positive region until the ratio satisfies the desired value γ . When the predefined γ value would expand the true negative outer boundary outside of the image area, these unavailable pixels are clipped and only those existing in the image area are used. The steps to produce the edge region, and subsequently the true negative region that satisfies the area ratio for irregularly shaped desired targets is shown in Algorithm 1.

Algorithm 1: Compute R_e to satisfy $\frac{|R_b|}{|R_a|} = \gamma$

```

Initialize  $\epsilon = 2\gamma\omega$ ;
 $R_e \leftarrow R_o.\text{buffer}(\epsilon)$ ;
while  $\frac{|R_b|}{|R_a|} > \gamma$  do
|  $\epsilon \leftarrow \epsilon - \delta$ ;
|  $R_e \leftarrow R_o.\text{buffer}(\epsilon)$ ;
end while
return  $R_e$ ;

```

The RandCrowns measure is developed by modifying each of the four terms of the RI to be appropriate for imprecisely labeled targets. The correct number of true positive pairs between a detected delineation and a desired labeled target is computed by the intersection of the delineation and the true positive region as shown in Equation 10,

$$a = |D \cap R_a|^2. \quad (10)$$

This term will promote delineations that completely enclose the true positive region of the desired target. If a delineation completely includes the true positive region then this true positive score will be maximized. Whereas if a delineation does not consist of the entire portion of the true positive region, the true positive score will be smaller.

RandCrowns ensures that not only do detected delineations encapsulate the true positive region, but also that delineations do not extend past the target encompassing background non-target regions. This is evaluated by computing the number of correct true negative pairs. The number of correct true negative pairs are computed as shown in Equation 11,

$$b = |R_b \setminus D|^2. \quad (11)$$

If the delineation does not surround portions of the true negative region, b will be high for a given delineation. If the delineation expands outside of the true negative inner boundary, b will decrease. Thus, b promotes delineations that do not extend outside the imprecise desired target. The false positive pairs and false negative pairs can also be computed.

To account for oversegmentations where the delineation contains regions outside of the desired crown area, the false positive pairs are counted. The false positive pairs are counted from the pixels within a delineation that cross into the true negative region as shown in Equation 12,

$$c = |D \cap R_b|^2. \quad (12)$$

If a delineation envelops portions outside of the true negative inner boundary then the false positive score will be high. Conversely, if the delineation only encloses portions within the true negative inner boundary, c will be low. This promotes small bounding boxes that do not include the true negative region surrounding the imprecisely labeled desired target.

Lastly, the number of false negative pairs are counted. These pixels are portions of the desired crown area that are missing from the delineation. The number of false negative pairs are pixels within the true positive region that are not within the delineation as shown in Equation 13,

$$d = |R_a \setminus D|^2. \quad (13)$$

The number of false negative pairs will be small if the delineation encompasses most of the true positive region. When smaller areas of the true positive region are enclosed within the delineation, the false negative score will be larger. Analogous to the Rand index, we combine the four terms defined above to create the RandCrowns index in Equation 14,

$$\text{RandCrowns} = \frac{a + b}{a + b + c + d}, \quad (14)$$

where the terms a, b, c , and d are defined for imprecise desired targets.

III. DATA

For this project we used orthorectified RGB photos collected from aircraft (0.1 m spatial resolution). The data was collected from the National Ecological Observatory Network Airborne Observatory Platform (NEON AOP) and clipped in four 40×40 m plots. Data was collected from three different forest ecosystems to better estimate the effect of forest type on crown annotation uncertainty and imprecision. Plots in the Ordway-Swisher Biological Station (OSBS) represent an open forest (with tree crowns relatively sparse); plots in the Mountain Lake Biological Station (MLBS) represent a generally closed forest (with crowns generally overlapping and more difficult to visually separate); plots in Talladega National Forest (TALL) represent a mix of open and closed forests. Two sets of annotations were used: one consists of polygons derived from field-measurements while the other was produced from remote sensing delineations.

The field polygons were generated by marking the location and boundary of individual tree crowns in NEON aerial photos while in the field [4]. Individual trees were located in the field plots using a GPS and the images for reference. The crown boundaries were digitized on the images to create the tree crown objects. Delineated crowns were refined in the lab using multiple NEON data products. These field delineations are available as part of a benchmark dataset for tree crown delineation [22]. The second set of annotations were labeled by four different experts using images alone, not in the field. Each was asked to hand delineate all visible crowns directly on each plot image. QGIS software was used to generate rectangular bounding boxes using a combination of RGB, hyperspectral and LiDAR data, following the approach proposed by [22]. Annotators were not allowed to view the annotations collected from the other three experts, to ensure independence and better estimate uncertainty in crown delineation among multiple annotators.

Due to the lack of a single true ground-truth due to variation among multiple annotations, we cross-

TABLE I
EXPERIMENT 1 DATASET. FOUR PLOTS FROM VARIOUS
GEOGRAPHIC LOCATIONS WERE LABELED BY FOUR
ANNOTATORS.

Plot	1	2	3	4	Total
Annotator 1	48	43	26	35	152
Annotator 2	48	54	31	58	191
Annotator 3	64	56	60	63	243
Annotator 4	70	33	34	54	191

validate delineations from four annotators where each annotator is considered the desired target and the other three are sample delineations. The delineations are from four diverse plots containing multiple trees per plot. Each plot is 40×40 m with 10 cm resolution from the NEON dataset. The total number of annotations for each annotator is listed in Table I.

The annotations were reduced to remove clearly mismatched annotations between the labelers (i.e., cases where the labels were for clearly different tree crowns). All annotations from each plot were visualized simultaneously by overlaying them on the RGB raster using QGIS and delineations of which there was little or no qualitative agreement among multiple annotators were removed. Within the raw data, multiple cases of unshared labels are present. In some instances, a subset of labelers did not label crowns that other annotators did. In other cases, one annotator labeled what visually appeared to be a single large crown with four separate canopies. However, other annotators labeled each of the four separate canopies as four individual crowns. Additionally, a pair of annotators did not label some small crowns that the other pair did, so these unshared delineations were also removed. In any case, where not all four annotators provided labels for the same crowns, these annotations were removed. We removed these annotations in order to rely on the performance of all possible annotators to assist for parameter selection. Without removing unshared delineations, each expert would not have the same weight in parameter selection. The mean crown size of Experiment 1 delineations was 39.6 m² and standard deviation 22.0 m². A

similar dataset was prepared given field-delineated annotations and two sets of annotations derived from the remote sensing data. The mean crown size of Experiment 2 field-delineations was 22.3 m² and standard deviation 15.8 m².

IV. EXPERIMENTS

To demonstrate our new RandCrowns method we provide two unique experiments. The first is a cross-validation experiment which compares multiple annotations of the same tree crowns delineated via remote sensing imagery by four annotators. Parameters for the model are modified to reduce the variance between annotators. While small variance between annotators is desired, qualitative results are considered influential to the selection of RandCrowns parameters. The second experiment applies the RandCrowns measure to a dataset consisting of field delineated crowns and scores bounding box annotations derived from the remote sensing data. Two annotators labeled the remote sensing data and the variance between their sample delineations is used to validate the RandCrowns measure.

A. Comparisons using multiple annotators

We quantitatively and qualitatively compared the RandCrowns and IoU scores based on labels generated using only imagery by multiple annotators. Due to the uncertainty in expert annotations, we do not assume a single set of “true” ground-truth labels that delineate the tree crowns in our dataset. Instead, we compare the labels created by the four annotators and measure the average variance between each set of three annotators based on treating the fourth as the desired target for comparison. We compute the average variance as,

$$\bar{S}^2 = \frac{1}{K} \sum_{k=1}^K \frac{\sum_{z=1}^Z (\chi(D_{z,k}, T_k) - \mu_{\chi_k})^2}{Z-1}, \quad (15)$$

where annotations from the three (Z) are treated as sample delineations ($D_{z,k}$), while the fourth is the desired target (T_k) in each experiment. There are K crowns in an experimental dataset and μ_{χ_k} is the average quantitative measurement (χ representing either IoU, IoUCrowns, or RandCrowns)

for the k th crown. As noted in Equation 15, each individual delineation and assigned target are given a quantitative score. We pair sample delineations to the desired target given the highest IoU score for each delineation-desired target pair, which leads to a conservative assessment of the new method. The main goal in developing the RandCrown metric is to achieve low variance among annotators, but given our target selection method, we also expect the metric to indicate a true delineation.

Given that the goal is to create a measure which assigns similar annotations values with minimal variance in the presence of uncertainty in labels, we computed the average variance of RandCrowns between annotators and varied the parameters of the RandCrowns measure. With the heuristic we established to compute the true negative outer boundary, there are three parameters to be chosen: the true positive boundary parameter, true negative inner boundary parameter, and the ratio of the true negative area to the true positive area (see Figure 1). We varied these parameters within two experiments. The first compared annotations from four annotators on tree crowns delineated from remote sensing imagery while the second compared annotations from two annotators on polygons derived from field measurements. The parameters selected for both experiments were chosen to: 1) minimize variance between annotators and 2) produce coherent positive and negative regions.

In addition to evaluating RandCrowns we also evaluated a buffered version of the IoU by disregarding true negatives in the RandCrowns calculations. Using buffers with the IoU instead of the Rand index resulted in higher variance between annotators than when relying on the RandCrowns measure (Table II). It also resulted in an increased variance even in comparison to using the IoU without buffers due to ignoring true negatives (Table II). This demonstrates the importance of considering true negatives in this context and so RandCrowns, not the modified IoU (IoUCrowns), was used for all further experiments.

To investigate the relationship between the IoU and RandCrowns scores we compared each pair of annotator measurements for each cross-validation

TABLE II
VARIATION OF RANDCROWNS AND IOU MEASURES ACROSS ANNOTATORS FOR THE CROSS-VALIDATION BOUNDING BOX EXPERIMENT.

Measure	Variance
RandCrowns	0.008
IoU	0.022
IoUCrowns	0.040

experiment given our chosen parameters (see Figure 2). Overall, the RandCrowns is more stable than the IoU with a variance roughly $\frac{1}{3}$ that of the IoU (Table II). While the RandCrowns assigns matching delineation-desired target pairs with high scores and low variance, the IoU does not. There is no direct relationship between the IoU and RandCrowns for the chosen parameters, but in general RandCrowns scores are higher than IoUs, low RandCrowns scores are associated with low IoU, and most RandCrowns scores are clustered near one.

In our experiments we were able to achieve zero variance between annotators using RandCrowns. However, the parameters used to obtain this agreement allowed delineations containing a large false positive region to be given a high score (see Figure 3). Therefore, we used the variance to guide parameter selection and achieved an error between the pair of annotators three times better than the IoU. This highlights the importance of considering uncertainty when comparing multiple annotators for tree crown delineation. Because the IoU only considers an intersection over union of the delineation and the desired target, there is no mechanism to address uncertainty and imprecision.

The parameter ranges for our experiments are shown in Table III. While there are an infinite number of combinations of parameters to try, we used the average variance between annotators in our cross-validation experiments to guide our choices along with qualitative evaluation. As previously mentioned, some parameter sets may lead to zero variance and we have included an example demonstrating this issue in Figure 3.

The true positive boundary parameter was varied between 10 cm and 1 m. This reflects an uncertainty

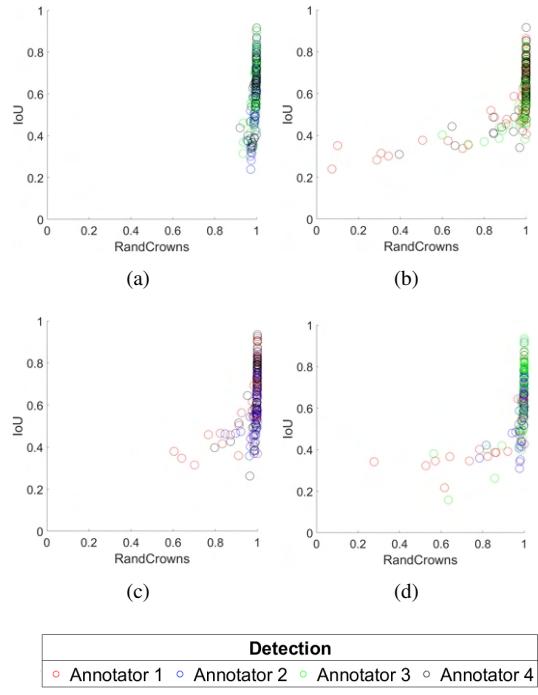


Fig. 2. The relationship between the RandCrowns score and the IoU measure for each experiment given our selected parameters. Each series represents an experiment of an annotator-annotator pair where the former is considered the delineation and the latter is the desired target. Each subfigure (a, b, c, and d) corresponds to the annotations from annotators 1-4 being the desired target respectively. Out of the four experiments, only a few examples consist of small RandCrowns scores. Some of these are shown in Figure 6.

of the desired target being 0.01 to 1 m^2 larger than the actual tree crown. The true negative inner boundary parameter was modified between 10 cm and 1.5 m; which reflects an uncertainty of the crown being 0.01 to 2.25 m^2 larger than the desired target. Lastly, the ratio of the true negative region to the true positive region (γ) was varied between 1 and 7. Both quantitatively minimal variance for the cross-validation experiments and qualitative relationships between the selected boundary sizes were used to select the parameters. The lowest average variance with reasonable boundary sizes, 0.008, was observed for true positive boundary size 0.7 m, true negative inner boundary size 1.2 m, and area ratio

TABLE III

RANDCROWNS PARAMETERS. EACH PARAMETER WAS VARIED AND THE AVERAGE VARIANCE ACROSS THE FOUR CROSS-VALIDATION EXPERIMENTS WAS RECORDED. THE NOTATION FOR THE PARAMETER RANGES ARE DEFINED TO BE $[l : s : u]$, WHERE l , s , AND u CORRESPOND TO THE LOWER BOUND, STEP SIZE, AND UPPER BOUND RESPECTIVELY. ADDITIONALLY, VISUALIZATIONS OF EACH SELECTED BOUNDARY SET WERE USED TO QUALITATIVELY ASSESS PERFORMANCE.

RandCrowns Parameters			
Parameter	α	ω	γ
Range	10 cm:10 cm:1 m	10 cm:10 cm:1.5 m	1:7

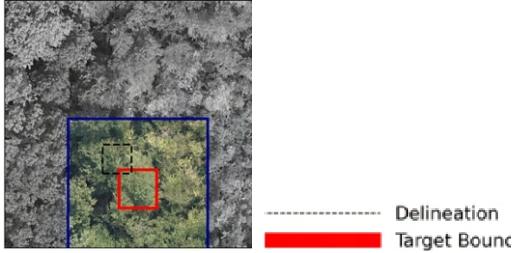


Fig. 3. An example of issues with missed delineations and inappropriate parameters. An example of a missed delineation that has an inappropriately high RC value ($RC = 0.99$, $IoU = 0.02$) given parameters that resulted in zero variance between annotators. The zero variance parameters are: true positive boundary 10 cm, true negative inner boundary 30 cm, and area ratio 7. The same delineation with our selected parameters; true positive boundary 70 cm, true negative inner boundary 1.2 m, and area ratio 3 scores a zero as a missed delineation ($RC = 0$, $IoU = 0.02$). As the true negative inner boundary parameter is too large with a slightly larger area ratio, all scores are near one even in cases where the delineations clearly do not match the desired target. It is possible that the high RandCrowns measure could be negated by reducing the area of the true positive boundary. However, that would merely fix this type of missed delineation and not necessarily be applicable for other delineations that would encompass small parts of the true positive boundary and much larger areas of an immense false region.

3. Although these are the selected parameters, it is possible to obtain a lower variance. If the area ratio is increased solely, the variance can be driven to zero. This case can be visualized in Figure 3.

Given an area ratio that is too large, even with a small true negative inner boundary parameter, the true negatives greatly outnumber the false positives. With the true negative region too large the true negatives greatly outweigh the false positives, resulting in some true positives counted in the delineation even though it is labeling another nearby tree. In order to avoid this issue it is necessary to set

reasonable limits on the range of parameters based on the details of the data that address the trade-off between absolute minimal variance and appropriate boundary sizes.

For the desired targets from comparing annotators 2 and 4 (Figure 2b and Figure 2d) there are multiple examples of delineations which score high with RandCrowns being scored below an IoU score of 0.5 which may be considered false delineations [23]; some of these interesting cases are shown in Figure 4.

In Figure 4a-d, the delineation is slightly larger than the desired target. While the IoU scores each of these with less than a 0.5 score, the RandCrowns measure assigns high scores greater than 0.9. For each of these example crown delineations, the fact that the desired target (i.e. the “labeled” crown) is larger than the delineation is not penalized, which is desirable because while the area of the crowns differ they both represent the same tree. This supports the importance of the outer region from the true negative inner boundary parameter. Due to the true negative inner boundary parameter being large for smaller crowns, the area between the desired target and the true negative inner boundary is scaled to allow larger delineations to achieve a good score. In Figure 4e-h, the four examples are each cases of smaller delineations than the desired target being granted low scores by the IoU. The IoU scores each of these delineations with less than 0.5 but the RandCrowns measure is greater than 0.96. This shows that although the delineations are not identical to the desired target, they each encompass a significant portion of the true positive boundary while intersecting none of the false delineation area.

While in some cases scores differ dramatically

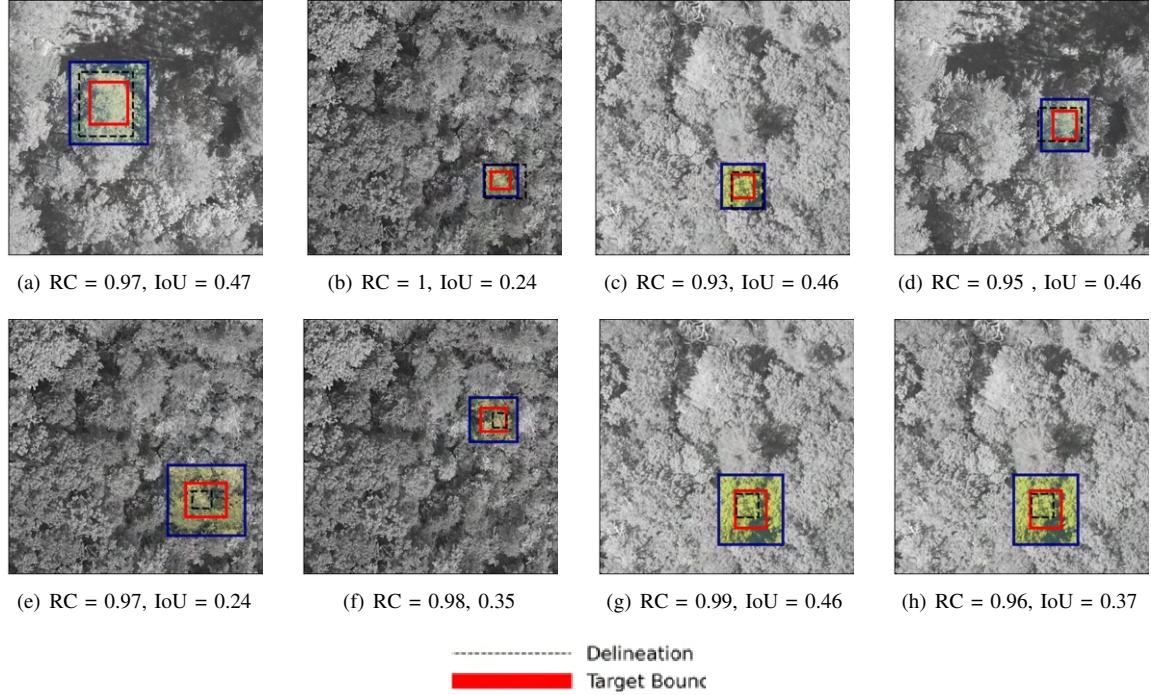


Fig. 4. Examples where the chosen delineation from the largest IoU value is given a much higher score by the RandCrowns measure. Inner and true negative inner boundaries are not plotted for clarity. Across a variety of desired target sizes, the RandCrowns assigns each of the delineations with a large score. Each case is an example of a delineation that would be considered a false positive given an IoU threshold of 0.5 from [23]. Given a smaller threshold of 0.4, half would be considered true positives [22].

between RandCrowns and IoU, in others the IoU and RandCrowns agree on a large score for the delineation. In cases where the IoU is > 0.6 the RandCrowns score is typically very high, close to one (Figure 2). Visual inspection of some example crowns in this range demonstrate the difficulty of achieving a high score with IoU unless the delineation is nearly identical to the desired target (Figure 5). This means that even small differences among annotators would result in meaningful differences in IoU even with effectively perfect delineation of the crown by an algorithm. Rather than requiring unrealistically perfect annotations, the RandCrowns measure grants a large score to delineations that differ slightly from the desired target but are qualitatively reasonable matches.

The RandCrowns measure also scores a variety of poor delineations in agreement with the IoU

appropriately (Figure 6). Examples where the IoU is below 0.4 show that the RandCrowns and IoU measures agree with low scores for the four delineation-desired target pairs. Although three of the four delineations shown in Figure 6 (a, b, and d) have higher RandCrowns scores, in the case of missed delineations the RandCrowns measure gives lower scores than the IoU as shown in Figure 7.

B. Comparing image-only annotations to field-delineated crowns

We apply the RandCrowns measure to an experiment comparing multiple annotators against field-delineated crowns. In this context we can consider the field delineated crown to be a perfect delineation (due to the higher accuracy of delineation conducted while in the field; [15]) occurring in the

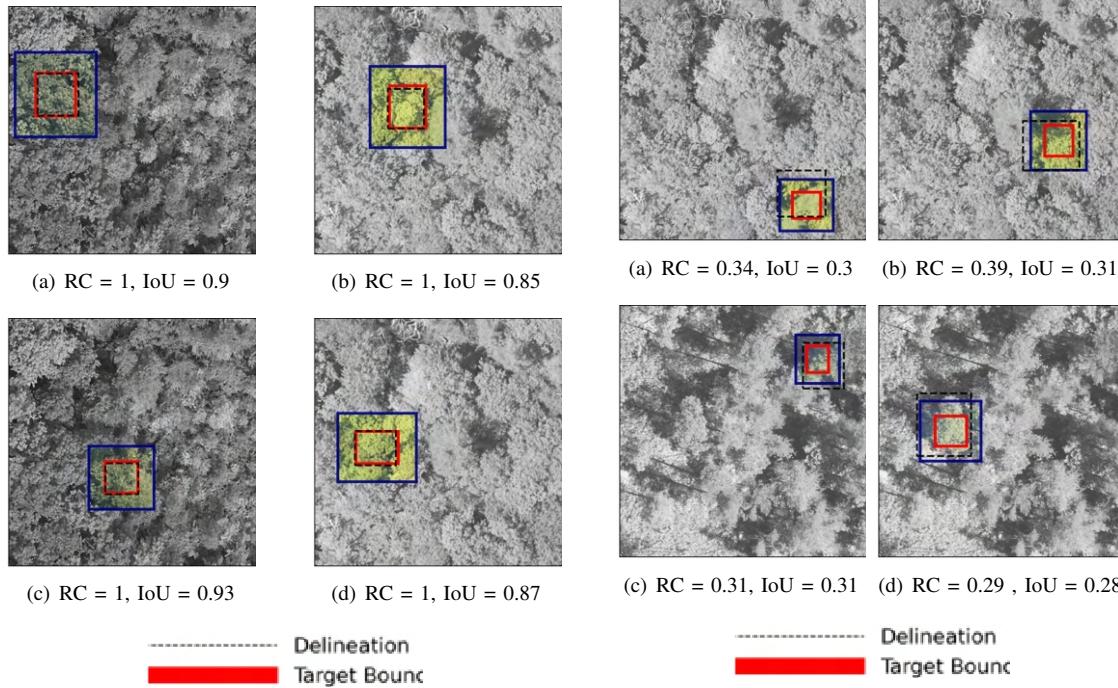


Fig. 5. Examples of some of the highest scoring IoU delineations from the cross-validation experiment. In each case the IoU and RandCrowns measure award a high score to the delineation because it is nearly identical to the desired target.

presence of uncertainty in the ground-truth labels (represented by the two image-only based delineations by different annotators). Our field dataset includes 22 polygons from 6 plots. The center point of each polygon was displayed to each annotator and the annotator labeled the crown surrounding the centroid with a bounding box to ensure that both annotators were labeling the same tree that had been delineated in the field. We repeat the procedure for the cross-validation experiment and select parameters that minimize the differences in scores between the two annotators. Due to the field-delineated crowns being smaller in this case than the image-based delineation, we found that using smaller true negative inner boundary sizes and area ratios provided less variation between annotators. The average variance between the pair of annotators for RandCrowns was approximately $\frac{1}{14}$ of

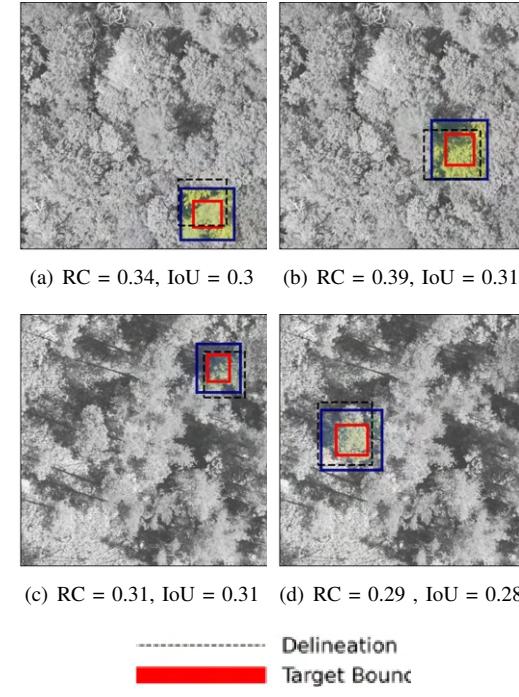


Fig. 6. Examples of low scoring delineation-desired target pairs where the RandCrowns and IoU agree on low scores.

TABLE IV
VARIATION OF RANDCROWNS AND IOU MEASURES ACROSS
ANNOTATORS FOR THE CROSS-VALIDATION BOUNDING BOX
VS. FIELD-DELINEATED POLYGONS EXPERIMENT.

Measure	Variance
RandCrowns	0.001
IoU	0.014

the variability for IoU (Table IV). Our parameters for the polygon implementation were: true positive boundary 60 cm, true negative inner boundary 3 m, and area ratio of 3.

Two out of the 22 comparisons ($\sim 10\%$) showed low IoU scores (< 0.5) associated with RandCrowns scores near one (Figure 8). In both cases the delineation scored poorly based on IoU because the area of the delineation is slightly larger than the desired target. If a threshold of 0.5 [23] was used to discard missed delineations, all four would be considered false positives. At most, one of these delineations

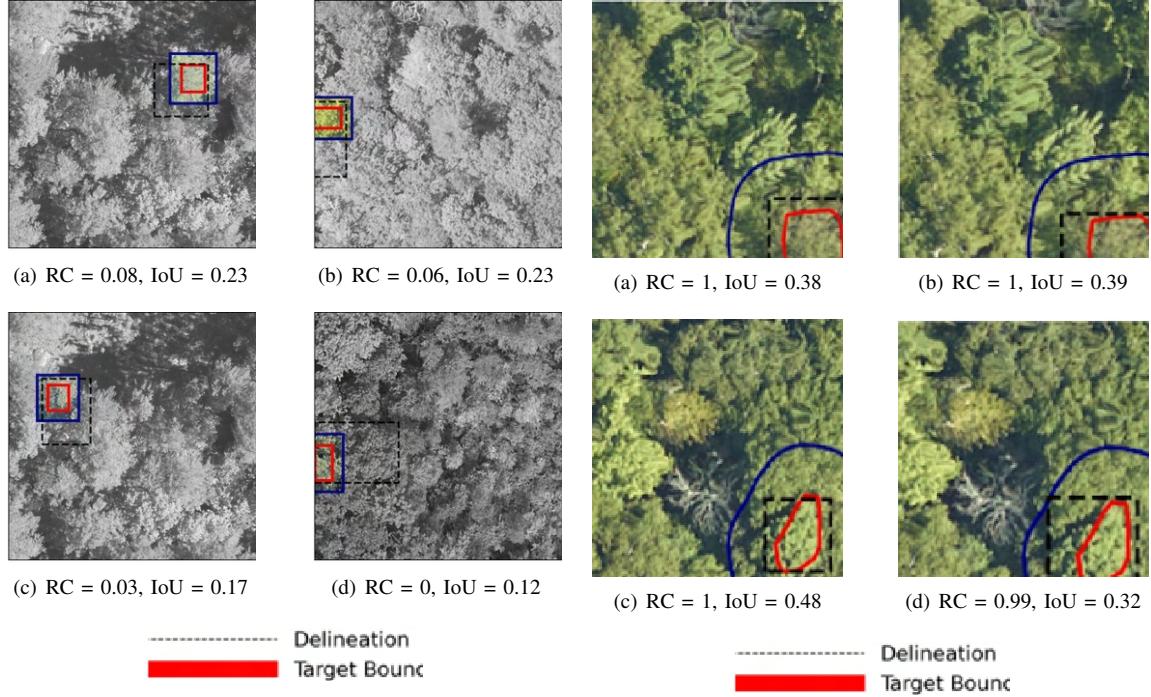


Fig. 7. Examples of missed delineations where the RandCrowns and IoU assign low scores. In each case the RandCrowns measure grants a lower score than the IoU.

would be considered a true positive given the 0.4 threshold recommended by [22]. In Figure 8a-b, the scores given by the IoU are almost identical for the two annotators. The same is true of the scores given by the RandCrowns. Meanwhile, in Figure 8c-d, for nearly equivalent delineations, the IoU has a 0.16 difference in score and the RandCrowns measure is consistent across the annotators. Rather than penalizing labels of the same tree that are not perfect replications of one another, the RandCrowns measure scores the delineations similarly across the annotators. Although the delineations of the two annotators are slightly larger than the desired target, each is labeling the desired crown. Therefore, a large score is warranted in each case and low difference (in this case zero) between the scores is appropriate.

Although we have demonstrated examples of applying the RandCrowns to pairs of delineations,

Fig. 8. Two comparisons between the annotators where the IoU grants low scores compared the RandCrowns measure. Annotator one labeled (a) and (c) while annotator two labeled (b) and (d). In three of the four cases: (a) (b) and (d), the IoU scores the delineation compared to the desired target with $\text{IoU} < 0.4$. Both crowns were labeled similarly from both bounding box annotators and achieved a zero difference between (a) and (b) and 0.01 between (c) and (d).

there are cases in application when a global quantitative score is desired. In order to obtain a global RandCrowns score there are many possibilities. To obtain a score for each target, a delineation needs to be assigned to every target. Delineations which do not match any target should reduce the overall global score. Matching a delineation to a target can be done by assigning the target the nearest delineation based on the minimum spatial distance between centers. However, it is possible that multiple delineations have the same distance to a target. In that case, we recommend returning the minimum due to observations from our preliminary results during development of the IDTreeS 2020

competition evaluation toolbox [24]. Relying on averaging scores produced by multiple delineations can promote poor delineation performance. Outside of our controlled experiments, if a delineation approach is applied which provides many delineations for each crown, there are multiple options to penalize the delineation method. Either the minimum score can be retained for each target, or the average can be kept and divided by the number of delineations assigned to the target crown. Another possible approach is to score every possible pair of delineation and target pairs and assign each target the candidate delineation with the maximum score. In the latter approach, delineations that are very large may be assigned to multiple crowns and obtain large scores for multiple crowns. We recommend using the first approach for assigning delineations to target crowns. Given a RandCrowns score for each target, the average and standard deviation of the RandCrowns score should be used to compare delineation methods.

In general, we recommend constructing a validation set consisting of multiple target delineations for each crown. There are several ways of producing multiple target delineations. As we have done in this study, multiple annotators can label each crown individually without knowledge of the other candidate labels. Additionally, target delineations could be proposed by performing warpings (e.g., scalings, rotations, translations) on each hand-labeled delineation. This approach should be investigated in future work. Given a set of target delineations for each crown, the RandCrowns parameters should be varied and those that return minimal average variance of RandCrowns score across the validation set should be proposed as possible parameter settings.

We demonstrate an application of our new RandCrowns measure to tree crown delineation and provide improvements over the existing IoU. However, several potential limitations should be discussed. Our parameter selection approach does not necessarily provide the optimal parameters for this problem. Certainly, there are other sets of parameter choices which would grant small variance between annotators. While we choose parameters that minimize variance between annotators, qualitative

checks are necessary to confirm. As mentioned in the previous paragraph, methods of synthetically generating reliable desired target delineations should be investigated in future work. Additionally, approaches to aggregate the RandCrowns score for a global measure should be tested. Our preliminary work revealed possible issues with assigning a candidate delineation to a target. We provide brief guidance on penalizing methods which produce an inappropriately large number of delineations or delineations of extreme size, however, there are more possibilities that should be studied.

V. CONCLUSION

This paper introduces the RandCrowns measure, an adaptation of the Rand index for weakly-labeled crown delineation. RandCrowns relies on new regions where each term of the Rand index is applied for imprecisely labeled desired targets containing uncertainty. A cross-validation experiment with four annotators was used to demonstrate the ability of the RandCrowns measure to reduce the variance in scores resulting from difference in human annotators with minimal average variance (0.008) in comparison to the commonly used IoU (0.022) and qualitative examples were provided to show the necessity of selecting useful parameters. The RandCrowns measure can also be applied to smoother polygons in addition to bounding boxes, as demonstrated by the second experiment. The parameters were selected to produce minimal variance between the two labelers and the variance was a factor of 14 better than the IoU. In this work we applied RandCrowns for tree crown delineation. However, RandCrowns can be used to quantify segmentation performance in any application that relies on imprecise labels.

ACKNOWLEDGMENT

Thanks to Ira Harmon and Daisy Wang for helpful discussions of evaluation methods related to tree crown delineation. This research was supported by the Gordon and Betty Moore Foundation’s Data-Driven Discovery Initiative through grant GBMF4563 to E.P. White and by the National

Science Foundation through grant 1926542 to E.P. White, S.A. Bohlman, A. Zare, D.Z. Wang, and A. Singh and grant 1442280 to S. A. Bohlman.

REFERENCES

- [1] S. Marconi, S. J. Graves, B. G. Weinstein, S. Bohlman, and E. P. White, "Estimating individual-level plant traits at scale," *Ecological Applications*, p. e02300, 2021.
- [2] M. Dalponte and D. A. Coomes, "Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data," *Methods in ecology and evolution*, vol. 7, no. 10, p. 1236–1245, October 2016. [Online]. Available: <https://europepmc.org/articles/PMC5137341>
- [3] FAO, "State of the world's forests 2016," 2016. [Online]. Available: <http://www.fao.org/publications/sofo/2016/en/>
- [4] S. J. Graves, T. T. Caughlin, G. P. Asner, and S. A. Bohlman, "A tree-based approach to biomass estimation from remote sensing data in a tropical agricultural landscape," *Remote Sensing of Environment*, vol. 218, pp. 32–43, 2018.
- [5] S. J. Graves, G. P. Asner, R. E. Martin, C. B. Anderson, M. S. Colgan, L. Kalantari, and S. A. Bohlman, "Tree species abundance predictions in a tropical agricultural landscape with a supervised classification model and imbalanced data," *Remote Sensing*, vol. 8, no. 2, 2016.
- [6] W. M. W. S. Jaafar, I. H. Woodhouse, C. A. Silva, H. Omar, K. N. Abdul Maulud, A. T. Hudak, C. Klauber, A. Cardil, and M. Mohan, "Improving individual tree crown delineation and attributes estimation of tropical forests using airborne lidar data," *Forests*, vol. 9, no. 12, 2018.
- [7] N. Saarinen, M. Vastaranta, R. Näsi, T. Rosnell, T. Hakala, E. Honkavaara, M. A. Wulder, V. Luoma, A. M. G. Tommaselli, N. N. Imai, E. A. W. Ribeiro, R. B. Guimaraes, M. Holopainen, and J. Hyppä, "Assessing biodiversity in boreal forests with uav-based photogrammetric point clouds and hyperspectral imaging," *Remote Sensing*, vol. 10, no. 2, 2018.
- [8] B. Hu, J. Li, L. Jing, and A. Judah, "Improving the efficiency and accuracy of individual tree crown delineation from high-density lidar data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 26, p. 145–155, 02 2014.
- [9] B. Wu, B. Yu, Q. Wu, Y. Huang, Z. Chen, and J. Wu, "Individual tree crown delineation using localized contour tree method and airborne lidar data in coniferous forests," *International Journal of Applied Earth Observation and Geoinformation*, vol. 52, pp. 82–94, 2016.
- [10] D. Mongus and B. Žalik, "An efficient approach to 3d single tree-crown delineation in lidar data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 108, pp. 219–233, 2015.
- [11] B. G. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, "Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks," *Remote Sensing*, vol. 11, no. 11, 2019.
- [12] S. Deng, M. Katoh, X. Yu, J. Hyppä, and T. Gao, "Comparison of tree species classifications at the individual tree level by combining als data and rgb images using different algorithms," *Remote Sensing*, vol. 8, no. 12, 2016.
- [13] L. Wang, P. Gong, and G. Biging, "Individual tree-crown delineation and treetop detection in high-spatial-resolution aerial imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 70, pp. 351–357, 03 2004.
- [14] R. Lucas, A. Lee, and M. Williams, "Enhanced simulation of radar backscatter from forests using lidar and optical data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 10, pp. 2736–2754, 2006.
- [15] S. Graves, J. Gearhart, T. T. Caughlin, and S. Bohlman, "A digital mapping method for linking high-resolution remote sensing images to individual tree crowns," *PeerJ Preprints*, 2018.
- [16] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bull. Soc. Vaud. Sci. Nat.*, vol. 44, pp. 223–270, 1908.
- [17] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.
- [18] D. Stewart, A. Hampton, A. Zare, J. Dale, and J. Keller, "The weakly-labeled rand index," 2021.
- [19] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, 2001, pp. 416–423 vol.2.
- [20] M. Meilă, "Comparing clusterings: An axiomatic view," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 577–584.
- [21] R. Unnikrishnan and M. Hebert, "Measures of similarity," in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, vol. 1, 2005, pp. 394–394.
- [22] B. G. Weinstein, S. J. Graves, S. Marconi, A. Singh, A. Zare, D. Stewart, S. A. Bohlman, and E. P. White, "A benchmark dataset for individual tree crown delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network," *bioRxiv*, 2020.
- [23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.
- [24] D. Stewart and S. Marconi, "weecology/idthrees_competition_evaluation: IDTREES competition 2020 evaluation code," May 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3833216>