Learning Neural Radiance Fields of Forest Structure for Scalable and Fine Monitoring

Juan Castorena[1]

Los Alamos National Laboratory, Los Alamos, NM, USA, 48124
`jcastorena@lanl.gov`

Abstract. This work leverages neural radiance fields and remote sensing for forestry applications. Here, we show neural radiance fields offer a wide range of possibilities to improve upon existing remote sensing methods in forest monitoring. We present experiments that demonstrate their potential to: (1) express fine features of forest 3D structure, (2) fuse available remote sensing modalities and (3), improve upon 3D structure derived forest metrics. Altogether, these properties make neural fields an attractive computational tool with great potential to further advance the scalability and accuracy of forest monitoring programs.

Keywords: Neural radiance fields · Remote Sensing · LiDAR · ALS · TLS · Photogrammetry · Forestry

1 Introduction

With approximately four billion hectares covering around 31% of the Earth's land area [7], forests play a vital role in our ecosystem. The increasing demand for tools that help maintain a balanced and healthy forest ecosystem is challenging due to the complex nature of various factors, including resilience against disease and fire, as well as overall forest health and biodiversity [25]. Active research focuses on the development of monitoring methods that synergistically collect comprehensive information about forest ecosystems and utilize it to analyze and generate predictive models of the characterizing factors. These methods should ideally be capable of effectively and efficiently cope with the dynamic changes over time and heterogeneity. The goal is to provide the tools with such properties for improved planning, management, analysis, and more effective decision-making processes [1]. Traditional tools for forest monitoring, such as national forest inventory (NFI) plots, utilize spatial sampling and estimation techniques to quantify forest cover, growing stock volume, biomass, carbon balance, and various tree metrics (e.g., diameter at breast height, crown width, height) [23]. However, these surveying methods consist of manual field sampling, which tends to introduce bias and poses challenges in terms of reproducibility. Moreover, this approach is economically costly and time-consuming, especially when dealing with large spatial extents.

Recent advancements, driven by the integration of remote sensing, geographic information and modern computational methods, have contributed to the development of more efficient, cost/time effective, and reproducible ecosystem characterizations. These advancements have unveiled the potential of highly refined and detailed models of 3D

forest structure. Traditionally, the metrics collected through standard forest inventory plot surveys have been utilized as critical inputs in applications in forest health [15], wood harvesting [13], habitat monitoring [24], and fire modeling [16]. The efficacy of these metrics relies in their ability to quantitatively represent the full forest's 3D structure including its vertical resolution: from the ground, sub-canopy to the canopy structure. Among the most popular remote sensing techniques, airborne LiDAR scanning (ALS) has gained widespread interest due to its ability to rapidly collect precise 3D structural information over large regional extents [6]. Airborne LiDAR, equipped with accurate position sensors like RTK (Real-Time Kinematic), enables large-scale mapping from high altitudes at spatial resolutions ranging from 5-20 points per square meter. It has proven effective in retrieving important factors in forest inventory plots [11]. However, it faces challenges in dense areas where the tree canopy obstructs the LiDAR signal, even with its advanced full-waveform-based technology. *In-situ* terrestrial laser scanning (TLS) on the other hand provides detailed vertical 3D resolution from the ground, sub-canopy and canopy structure informing about individual trees, shrubs, ground surface, and near-ground vegetation at even higher spatial resolutions [10]. Recent work by [20] has demonstrated the efficiency and efficacy of ecosystem monitoring using single scan in-situ TLS. The technological advances of such models include new capabilities for rapidly extracting highly detailed quantifiable predictions of vegetation attributes and treatment effects in near surface, sub-canopy and canopy composition. However, these models have only been deployed across spatial domains of a few tens of meters in radius due to the existing inherited limitations of TLS spatial coverage [20]. On the other side of the spectrum, image based photogrammetry for 3D structure extraction offers the potential of being both scalable and the most cost efficient. Existing computational methods for the extraction of 3D structure in forest ecosystems, however, have not been as efficient. Aerial photogrammetry methods result in 3D structure that contains very limited structural information along the vertical dimension and have encountered output spatial resolutions that can be at most only on par with those from ALS [25].

Our contribution seeks to fuse the experimental findings across remote sensing domains in forestry; from broad-scale to in-situ sensing sources. The goal is the ability to achieve the performance quality of *in-situ* sources (e.g., TLS) in the extraction of 3D forest structure at the scalability of broad sources (e.g., ALS, aerial-imagery). We propose the use of neural radiance field (NERF) representations [17] which account for the origin and direction of radiance to determine highly detailed 3D structure via view-consistency. We observe that such representations enable both the fine description of forest 3D structure and also the fusion of multi-view multi-modal sensing sources. Demonstrated experiments on real multi-view RGB imagery, ALS and TLS validate the fine resolution capabilities of such representations as applied to forests. In addition, the performance found in our experiments of 3D structure derived forest factor metrics demonstrate the potential of neural fields to improve upon the existing forest monitoring programs. To the best of our knowledge, the demonstrations conducted in this research, namely, the application of neural fields for 3D sensing in forestry, is novel and has not been shown previously. In the following, Sec. 2 provides a brief overview of neural fields. Sec. 3 includes experiments illustrating the feasibility of neural fields to

represent fine 3D structure of forestry while Section 4 demonstrates the effectiveness of fusing NERF with LiDAR data by enforcing LiDAR point cloud priors. Finally, Section 5 presents results that show the efficacy of NERF extracted 3D structure for deriving forest factor metrics, which are of prime significance to forest managers for monitoring.

2   Background

2.1   Neural Radiance Fields

The idea of neural radiance fields (NERF) is based on classical ray tracing of volume densities [12]. Under this framework, each pixel comprising an image is represented by a ray of light casted onto the scene. The ray of light is described by $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with origin $\mathbf{o} \in \mathbb{R}^3$, unit ℓ_2-norm length direction $\mathbf{d} \in \mathbb{R}^3$ (i.e., $\|\mathbf{d}\|_2 = 1$) and independent variable $t \in \mathbb{R}$ representing a relative distance. The parameters of each ray can be computed through the camera intrinsic matrix $\mathbf{K}$ with inverse $\mathbf{K}^{-1}$, the 6D pose transformation matrix $\mathbf{T}_{m \to 0}$ of image m as in Eq. (1)

$$(\mathbf{o}, \mathbf{d}) = \left(T_{m \to 0}^{(4)}, \frac{\mathbf{d}'}{\|\mathbf{d}'\|_{\ell_2}} \right) \quad \text{with}$$

$$\mathbf{d}' = \mathbf{T}_{m \to 0}^{-1} \mathbf{K}^{-1} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} - T_{m \to 0}^{(4)} \tag{1}$$

where u', v' are vertical and horizontal the pixel locations within the image and the subscript $^{(i)}$ denotes the i-th column of a matrix. Casting rays $\mathbf{r} \in \mathcal{R}$ into the scene from all pixels across all multi-view images provides information of intersecting rays that can be exploited to infer 3D scene structure. Such information consists on sampling along a ray at distance samples $\{t_i\}_{i=1}^M$ and determine at each sample if the color $\mathbf{c}_i \in [0, .., 255]^3$ of the ray coincides with those from overlapping rays. If it does not coincide then it is likely that the medium found at that specific distance sample is transparent whereas the opposite means an opaque medium is present. With such information, compositing color can be expressed as a function of ray $\mathbf{r}$ as in Eq. (2) by:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^N \left[\underbrace{\left(\prod_{j=1}^{i-1} \exp(-\sigma_j \delta_j) \right)}_{\text{transparency so far}} \underbrace{(1 - \exp(-\sigma_i \delta_i))}_{\text{opacity}} \mathbf{c}_i \right] \tag{2}$$

where $\sigma_i \in \mathbb{R}$ and $\delta_i = t_{i+1} - t_i$ are the volume densities and differential time steps at sample indexed by i, respectively. In Eq. (2) the first term in the summation represents the transparent samples so far while the second term is an opaque medium of color $\mathbf{c}_i$ present at sample i. Reconstructing a scene in 3D can then be posed as the problem of finding the sample locations t_i where each ray intersects an opaque medium (i.e., where each ray stops) for all rays casted into the scene. Those intersections are likely to occur at the sample locations where the volume densities are maximized; in other words,

where $t_i = \arg\max_i\{\sigma\}$. Accumulating, all rays casted into the scene and estimating the locations t_i's where volume density is maximized overall rays, renders the 3D geometry of the scene. The number of rays required per scene is an open question; the interested reader can go to [3] where a similar problem but for LiDAR sensing determines the number of pulses required for 3D reconstruction depending on a quantifiable measure of scene complexity.

The problem in Eq. (2) is solved by learning the volume densities that best explains image pixel color in a 3D consistent way. Learning can be done through a multilayer perceptron (MLP) by rewriting Eq. (2) as in Eq. (3) as:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} \mathbf{w}_i \mathbf{c}_i \tag{3}$$

where the weights $\mathbf{w} \in \mathbb{R}^N$ encode transparency or opacity of the N samples along a ray and $\mathbf{c}_i$ is its associated pixel color. Learning weights is performed in an unsupervised fashion through the optimization of a loss function using a training set of M pairs of multi-view RGB images and its corresponding 6D poses $\{(\mathbf{y}_m, \mathbf{T}_m)\}_{m=1}^{M}$, respectively. This loss function $f : \mathbb{R}^L \to \mathbb{R}$ is the average ℓ_2-norm error between ground truth color and estimation by compositing described as in Eq. (4):

$$\mathcal{L}_C(\mathbf{\Theta}) = \sum_{\mathbf{r} \in \mathcal{R}} \left[\|C(\mathbf{r}) - \hat{C}(\mathbf{r}, \mathbf{\Theta})\|_{\ell_2}^2 \right] \tag{4}$$

Optimization by back-propagation yields the weights that gradually improves upon the estimation of the volume densities. Other important parameters of NERF are distance $\hat{z}(\mathbf{r})$ which can be defined using the same weights from Eq.(2) but here expressed in terms of distance as:

$$\hat{z}(\mathbf{r}) = \sum_{i=1}^{N} \omega_i t_i, \qquad \hat{s}(\mathbf{r})^2 = \sum_{i=1}^{N} \omega_i (t_i - \hat{z}(\mathbf{r}))^2 \tag{5}$$

and $\hat{s}(\mathbf{r})$ defined as the standard deviation of distance. One key issue affecting 3D reconstruction resolution is on the way samples $\{t_i\}_{i=1}^{N}$ for each ray $\mathbf{r} \in \mathcal{R}$ are drawn. A small number of samples N results in low resolution and erroneous ray intersection estimations while sampling vastly results in much higher computational complexities. To balance this trade-off, the work in [17] uses two networks one at low-resolution to coarsely sample the 3D scene and another fine-resolution one used subsequently to more finely sample only at locations likely containing the scene.

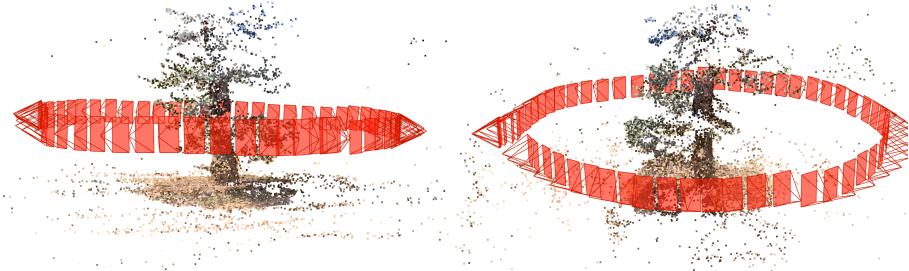## 3   Are neural fields capable of extracting 3D structure in forestry?

The high capacity of deep learning (DL) models to express data distributions with high fidelity and diversity offers a promising avenue to model heterogenous 3D forest structures in fine detail. The specific configuration of the selected DL model aims to provide a representation that naturally allows the combination of data from multiple sensing modalities and view-points. Neural fields [17] under the DL rubric have proven to be a highly effective computational approach for addressing such problems. However, their application has been only demonstrated for indoor and urban environments.

3.1 Terrestrial Imagery.

Expanding on the findings of neural fields in man-made environments, we conducted additional experiments to demonstrate its effectiveness in representing fine 3D structure details in forest ecosystems. Figs. 1 and 2 shows the extracted 3D structure of a Ponderosa pine tree in New Mexico, captured using standard 12-megapixel camera phone images collected along an elliptical trajectory around the tree. Fig. 1a shows a few of the input example terrestrial multi-view RGB images collected. Figs. 1b and 1c presents the image snapshot trajectory represented as red rectangles, along with two 3D structure views derived from a traditional structure from motion (SFM) method [22] applied to the multi-view input images. Note that the level of spatial variability detail provided by this SFM method is significantly low considering the resolution provided by the set of input images.

(a) Terrestrial RGB multi-view imagery of Ponderosa Pine Tree.

(b) SFM reconstruction view-1   (c) SFM reconstruction view-2

Fig. 1: Even though SFM reconstruction is capable of extracting the 3D structure of tree, its recontruction suffers from sparsity. Such sparsity limits the spatial variability of structure that can be captured thorugh such models.

Can the representational power of modern AI models do better than classical 3D structure extraction methods in Forestry? We extract 3D structure by neural fields using the same input images and obtain the result shown in Fig.2. Note that much finer spatial variability details can be resolved across the 3D structure including the ground, trunk, branches, leaves. Even fine woody debris as shown in Figs2c-.2d and, bark can be re-

solved as shown in Figs.2e-2f in contrast to the result of traditional SFM in Fig.1. Note that even points coming from images degraded by sun-glare as shown in Fig.1a landed in the tree within reasonable distances as shown in Fig.2a, this is significant specially considering the severity of the glare effects present in the 2D RGB images. In general, terrestrial multi-view imagery based NERF can be used to extract fine 3D spatial resolution along the vertical dimension of a tree stand with a level of detail similar to TLS and with the additional advantage of providing color for every 3D point estimate.

(a) Side views illustrating high 3D spatial detail along the vertical tree stem

(b) Tree 3D structure view-5   (c) Fine 3D resolution of forest floor structure

(d) Forest floor 3D structure   (e) Tree trunk view-1   (f) Tree trunk view-2
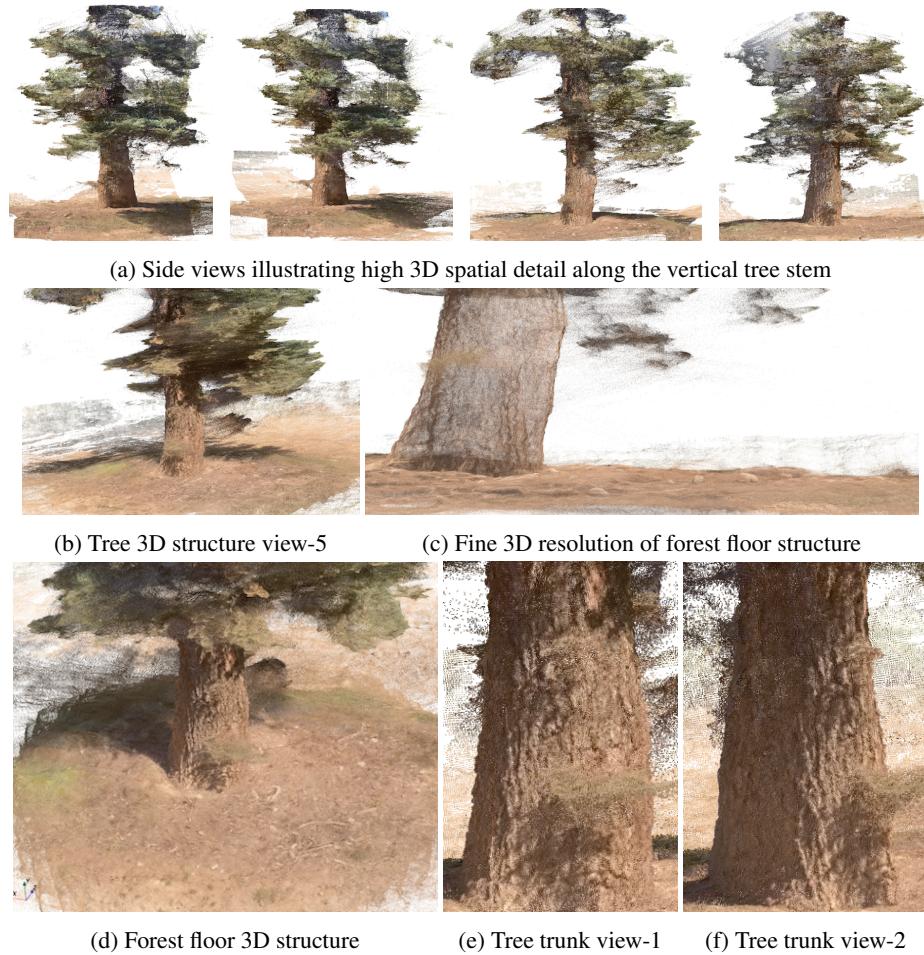
Fig. 2: Neural field models are capable of extracting fine 3D structure from terrestrial multi-view images in forestry. Reconstructions demonstrate their potential to represent fine scale variability in heterogeneous forest ecosystems.

4 Neural Radiance Fields: A framework for remote sensing fusion in forestry

Neural fields, have also demonstrated their ability to provide representations suitable for combining data from multiple sensing modalities in as long as these are co-registered or aligned. The neural fields framework, which extracts 3D structure from multi-view images, enables direct fusion of information with 3D point cloud sources through point cloud prior constraints [21]. Here, we consider the case of fusing multi-view images from an RGB camera and point clouds from LiDAR. The difficulty in fusing camera and LiDAR information is that camera measures color radiance while LiDAR measures distance [5]. Fortunately, the framework of neural radiance fields can be used to extract 3D structure from images thus enabling direct fusion of information from LiDAR. This can be done though a learning function that extracts a 3D structure promoting consistency between the multi-view images as leveraged by standard NERF [17] subject to LiDAR point cloud priors [21] as:

$$\mathcal{L}(\boldsymbol{\Theta}) = \underbrace{\sum_{\mathbf{r} \in \mathcal{R}} \left[\|C(\mathbf{r}) - \hat{C}(\mathbf{r}, \boldsymbol{\Theta})\|_{\ell_2}^2 \right]}_{\mathcal{L}_C(\boldsymbol{\Theta})} + \lambda \underbrace{\sum_{\mathbf{r} \in \mathcal{R}} \left[\|z(\mathbf{r}) - \hat{z}(\mathbf{r}, \boldsymbol{\Theta})\|_{\ell_2}^2 \right]}_{\mathcal{L}_D(\boldsymbol{\Theta})} \tag{6}$$

where the first term $\mathcal{L}_C(\boldsymbol{\Theta})$ is the standard NERF learning function promoting a 3D structure with consistency between image views while the second term $\mathcal{L}_D(\boldsymbol{\Theta})$ enforces the LiDAR point cloud priors with $\hat{z}(\mathbf{r}, \boldsymbol{\Theta})$ given as in Eq.(5). The benefit of imposing point cloud priors into neural fields is two-fold: (1) it enables expressing relative distances obtained from standard 3D reconstruction of multi-view 2D images in terms of real metrics (e.g., meters), and (2) neural fields tend to face challenges in accurately estimating 3D structures at high distances (typically in the order of several tens of meters), where the LiDAR point cloud priors can serve as a supervisory signal to guide accurate estimation, especially at greater distances. This can be beneficial, as distances in aerial imagery are generally distributed around large distances, which may pose challenges for 3D structure extraction methods.

4.1 Filing in the missing below-canopy structure in ALS data with TLS

In-situ terrestrial laser scanning (TLS) has been demonstrated as a powerful tool for rapid assessment of forest structure in ecosystem monitoring and characterization. It is capable of very fine resolution including the vertical direction: surface, sub-canopy and canopy structure. However, its utility and application is restricted by limited spatial coverage. Aerial laser scanning (ALS) on the other hand, has the ability to rapidly survey broad scale areas at the landscape level, but is limited as it sparsely samples the scene providing only coarse spatial variability details and it also cannot penetrate the tree canopy. Fig. 3a shows a point cloud example collected using a full-waveform ALS system which collects ≈ 10 points per meter square. In Fig. 3a note that the sub-canopy structure is not spatially resolved. In contrast, TLS is finely resolved below the canopy as observed in Fig.3b.
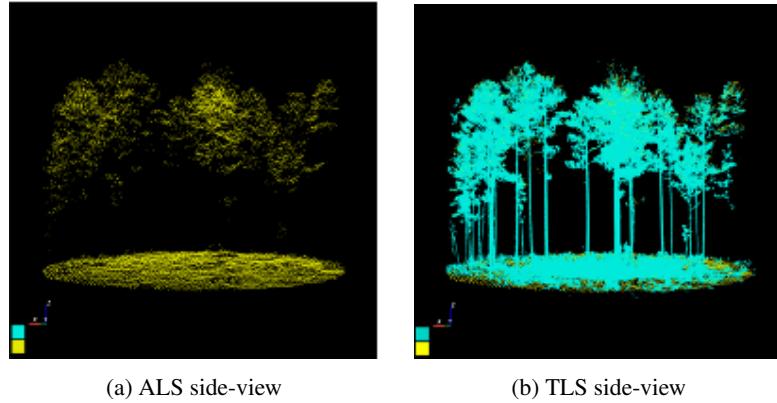
(a) ALS side-view   (b) TLS side-view

Fig. 3: Forest structure from TLS and ALS: ALS provides sparse spatial information and is not capable of resolving sub-canopy detail. TLS on the other hand, provides fine spatial variability and resolution along full 3D vertical stands.

Fortunately, the drawbacks of TLS and ALS scans can be resolved by co-registration which transforms the data to enable direct fusion. Here, we use the automatic and target-less based approach of [4]. This was demonstrated to outperform standard methods [2], [19], [8] in natural ecosystems and to be robust to resolution scales, view-points, scan area overlap, vegetation heterogeneity, topography and to ecosystem changes induced by pre/post low-intensity fire effects. It is also fully automatic, capable of self-correcting in cases of noisy GPS measurements and does not require any manually placed targets [9] while performing at the same levels of accuracy. All TLS scans where co-registered into the coordinate system of ALS. Once scans have been co-registered they can be projected into a common coordinate system. Illustrative example results for two forest plots where included in Fig.4 where the two sources: ALS and TLS have been color coded differently, with the sparser point cloud being that of the ALS. Throughout all cases the co-registration produced finely aligned point clouds. In general, the error produced by this co-registration method is <6 cm for the translation and $<0.1º$ for the rotation parameters. The translation error in mainly due to the resolution of ALS at 10 points/meter square.

4.2   Aerial Imagery

Experiments performed on broader forest areas were also conducted. Aerial RGB imagery was collected with a DJI Mavic2 Pro drone at 30Hz and a 3840×2160 pixel resolution. Figs. 5a-5f show examples of multi-view aerial image inputs used by the SFM and neural fields models. The forest 3D structure resulting from running conventional SFM [22] on these images is in Figs.5i-5k illustrating different perspective views. Again, the sequence of rectangles in red illustrate the drone flight path and the snapshot image locations. Note that SFM was capable of resolving 3D structure for the entire scene.

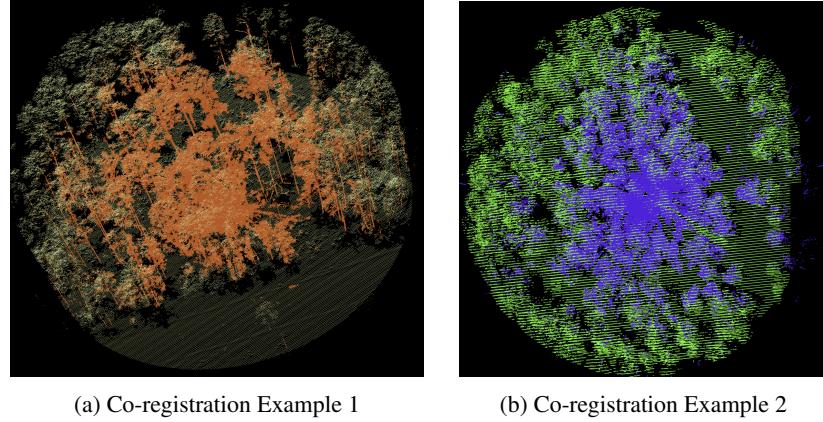(a) Co-registration Example 1   (b) Co-registration Example 2

Fig. 4: TLS to ALS co-registration: Forest features are well aligned qualitatively between both ALS and TLS sensing.

Applying NERF directly into the RGB imagery dataset, did not result in comparable performance as in the case of the Ponderosa pine tree shown in Section 3.1. Without point cloud constraints, the 3D structure extracted by the neural fields in Fig. 5h shows the presence of artifacts at large distances. The main reason for these artifacts is that NERF had difficulties in recovering 3D structures from images with objects distributed at far distances (e.g., ground surface in aerial scanning). Imposing LiDAR point cloud priors we hypothesize can help to alleviate this issue. Here, we follow the methodology of [21] and conduct experiments for fusing camera and LiDAR information through the learning function in Eq.(6). The LiDAR point cloud uses both co-registered TLS and ALS data which provides information to constrain both distances in the mid-story below the canopy and those between the ground surface and the tree canopy. The co-registration approach used to align ALS and TLS point clouds is the one described in Section 4.1. Note that TLS information is not available throughout the entire tested forest area; rather, only one TLS scan was collected. We found the information provided by just one single scan was enough to constraint the relative distances in sub-canopy areas throughout the entire scene. Imposing additional constraints through consistency with the input point cloud shown in Fig.5g, results in the extracted 3D structure shown in Figs. 5l-5n. In this case, the point cloud prior imposes constraints that resolve the associated difficulties at large distances. Note that this reconstruction is significantly less sparser than those shown in Figs.5i-5j obtained from conventional SFM. NERF+LIDAR results in improved resolution which in turn enables the detection of a much finer spatial variability, specially important for current existing demands in forest monitoring at broad scale. This illustrates the capacity of neural fields models not only to represent highly detailed 3D forest structure from aerial multi-view data but also the possibility of combining multi-source remotely sensed data (i.e., imagery and LiDAR).

(a) Image view-1   (b) Image view-2

(c) Image view-3   (d) Image view-4

(e) Image view-5   (f) Image view-6   (g) Point Cloud   (h) NERF artifacts

(i) COLMAP view-1   (j) COLMAP view-2   (k) COLMAP view-3

(m) NERF+LIDAR view-2

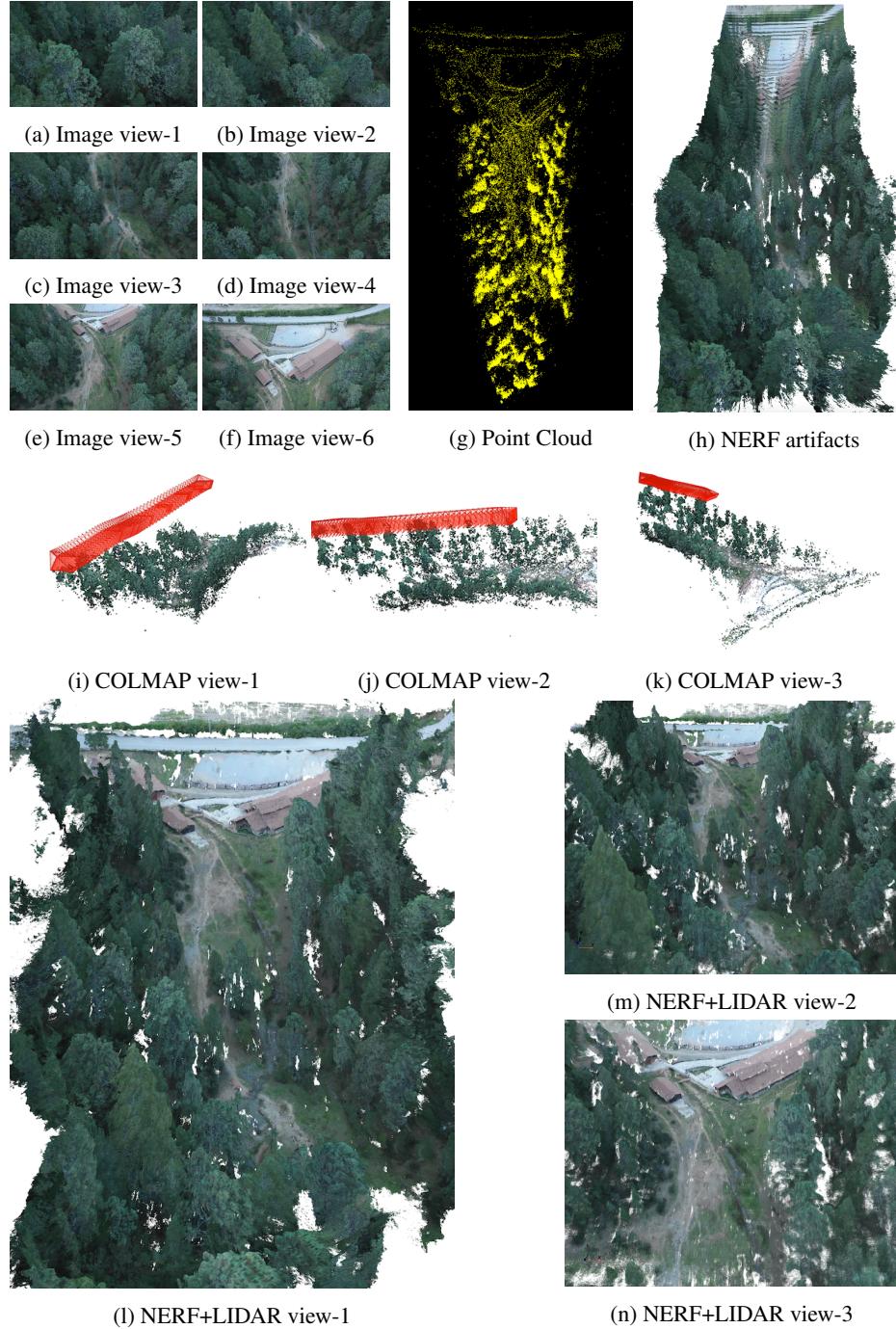(l) NERF+LIDAR view-1   (n) NERF+LIDAR view-3

Fig. 5: AI-based extraction of 3D structures from aerial multi-view 2D images + 3D point cloud data inputs. Imposing point cloud priors into 3D structure extraction improves distance ambiguities in structure and resolves artifact issues likely at far ranges.

5   Prediction of forest factor metrics

Demonstration of the described capabilities of neural fields on forest monitoring programs consists here in performance evaluations of 3D forest structure derived metrics. These can include for example number of trees, species composition, tree height, diameter at breast height (DBH), age on a given geo-referenced area. However, since our focus is to demonstrate the usefulness of neural radiance fields for representing 3D forest structure, we only illustrate its potential in prediction of the number of trees and DBH along geo-referenced areas. The data used includes overlapping TLS+ALS+GPS+aerial imagery multi-view multi-modal data collected over forest plot units. Each of these plots represents a location area of a varying size: some of size 20 x 50 m and others at 15 m radius. The sites in which data was collected is in northern New Mexico, USA (the NM dataset). The vegetation heterogeneity and topography variability of the landscape is significantly diverse. The NM site contains high elevation ponderosa pine and mixed-conifer forest: white fir, limber pine, aspen, Douglas fir and Gambel oak and topography is at high elevation and of high-variation (between 5,000-10,200 ft). The TLS data was collected using a LiDAR sensor mounted on a static tripod placed at the center of each plot. The ALS data was collected by a Galaxy T2000 LiDAR sensor mounted on a fixed-wing aircraft. The number of LiDAR point returns per volume depend on the sensor and scanning protocol settings (e.g., TLS or ALS, range distribution, number of scans) and these vary across plots depending on the heterogeneity of the site. Ground truth number of trees per plot was obtained through standard forest plot field surveying techniques involving actual physical measurements of live/dead vegetation composition. Data from a total of 250 plots where collected in the NM dataset. In every forest plot overlapping ALS, GPS, TLS and multi-view aerial imagery data was collected along with the corresponding field measured ground truth. Prediction of the number of trees y_1 per plot given point cloud $\mathbf{X}$, was performed following the approach of the GR-Net [27] [26]. In general, the methodology consists in computing 2D bounding boxes each corresponding to a tree detection from a birds eye view (BEV). A refinement segmentation approach then follows which projects each 2D bounding box into 3D space. The resulting points inside each 3D bounding box are then segmented by foliage, upper stem and lower stem and empty space and this information is used to improve estimates over the number of trees. This methodology is used independently on several case scenarios comparing the performance of a combination of remote sensing approaches: (1) neural fields (NF) from aerial RGB Images, (2) ALS as in Fig.6b, (3) TLS as in Fig.6a, (4) ALS+TLS, (5) NF-RGB images + ALS, (6) NF-RGB images + TLS, (7) NF-RGB Images + TLS + ALS. Note that the TLS, ALS and TLS+ALS prediction results does not make any use of neural fields. Rather, their performance was included only for comparison purposes. Table 1 summarizes the root mean squared error (RMSE) results for each of the tested cases.

The results in Table 1 corroborate some of the trade-offs between the sensing modalities and in addition some of the advantages gained through the use of neural fields in forestry. First, the superiority of TLS over ALS data on the number of trees metric is mainly due to the presence of information in sub-canopy which is characteristic of in-situ TLS. This in alignment with current demonstrations in the literature which have motivated the widespread usage of in-situ TLS in forestry applications even though

(a) In-situ plot-scale TLS has demonstrated to be an effective tool in estimating plot-level vegetation characteristics

(b) Broad-landscape scale ALS derived prediction, does not have vertical dimension resolution resulting in underestimate predictions
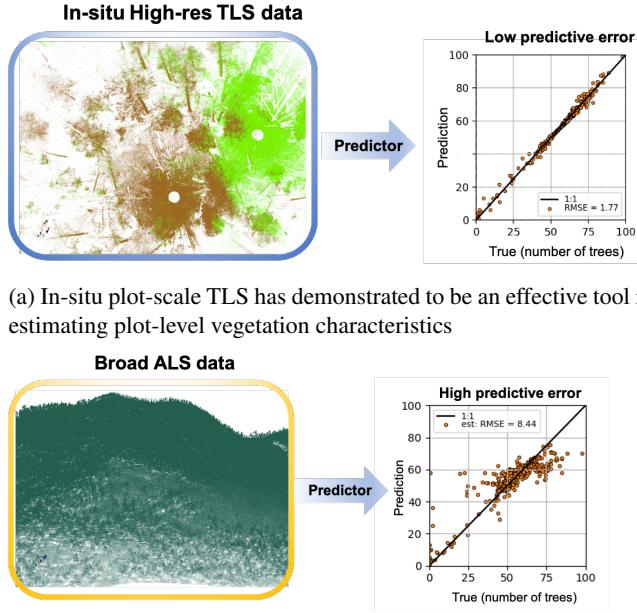
Fig. 6: LiDAR derived vegetation attribute estimation for single TLS and ALS.

Table 1: RMSE Prediction performance of number of trees per plot in NM dataset.

Method	NF-RGB	ALS	TLS	ALS+TLS	NF-RGB+ALS	NF-RGB+TLS	NF-RGB+ALS+TLS
RMSE	10.61	8.44	1.77	1.67	1.41	1.39	1.32

it is not as spatially scalable as ALS is [20]. We would have seen the opposite relationships between TLS and ALS, however, in cases when the plot size is significantly higher than the range of a single in-situ TLS scan. A problem which can be resolved by adding multiple view co-registered TLS scans per plot. This limitation is caused as the sensor remains static at collection time which makes it more susceptible to occlusions, specially in dense forest areas where trees can significantly reduce the view of TLS at higher ranges. TLS+ALS overcomes, on the other hand, the limitations of the individual LiDAR platforms by filling in the missing information characteristic of each platform. Structure from neural fields using only multi-view RGB images performed slightly worst than both ALS and TLS. This may be due to the limited number of multi-view images collected per plot, the performance for deriving structure from NERF or to the joint performance of NERF in conjunction with the GRNet. Fortunately, fusing neural fields from multi-view imagery with LiDAR shows a significant improvement overall fused cases (i.e., NF+ALS, NF+TLS and NF+ALS+TLS). We see that the prior

supervisory signal imposed by the LiDAR point cloud helps on guiding the resulting 3D structure from NERF to alleviate the artifacts arising at far distances when using multi-view imagery only. We would like to finalize this discussion by highlighting the performance of the NF-RGB+ALS method which is marginally similar to the best performing method (i.e., NF-RGB+ALS+TLS). The benefit of using NF-RGB+ALS is that being both airborne makes the data collection of these two modalities time and cost efficient, in contrast, to in-situ remote sensing methods such as TLS. This has significant implications towards achieving both scalable and highly performing forest monitoring programs. In general, one has to resort to a balance between scalability and performance performance depending on needs. Our work instead, offers a method which can potentially achieve similar performance as in-situ methods with the benefits of scalability over the landscape scales through computational methods.

Additional experiments were conducted to explore the ability of neural fields from terrestrial based multi-view imagery to achieve a performance near that of TLS in metrics that depend on sub-canopy information. In this case, we evaluated performance on the DBH metric for a total of 200 trees. Ground truth DBH was manually measured in the field for each tree's stem diameter at a height of 1.3m. A total of 5 co-registered TLS scans where used per tree, each collected from a different location and viewing each tree from a different perspective to reduce the effects of occlusion and to remove the degrading effects of lower point LiDAR return densities at farther ranges. Multi-view TLS co-registration was obtained using the method of [4]. Terrestrial multi-view RGB imagery data for NERF was collected around an oblique trajectory around each tree as exemplified in Fig.1 with $10 - 15$ snapshot images per tree. Algorithmic performance for estimating DBH was compared against TLS, ALS, TLS+ALS and NF-RGB. The estimation approach of [26] relying on stem geometric circular shape fitting at a height of 1.3m over the ground was used following their implementation. Performance is measured as the average error as a percentage of the actual field measured DBH ground truth, following the work of [26]. Comparison results are reported in Table 2.

Table 2: Comparison of sensing modalities on average error DBH estimation.

Method	NF-RGB	ALS	TLS	ALS+TLS
Avg. error %	1.7 %	32.7%	1.3%	3.3%

In table 2 ALS performs the worst DBH estimation due to its inherited limited sub-canopy resolution. Multi-view TLS on the other hand, performs the best at 1.3% error consistent with TLS superiority findings in [26] for metrics relying on sub-canopy information. However, our neural fields approach from terrestrial imagery performs marginally on par with multi-view TLS, with the additional advantage that RGB camera sensors are simpler to access commercially and significantly cheaper than LiDAR.

In terms of computational specifications, neural radiance fields were trained using a set of overlapping 10-50 multi-view images per scene. The fast implementation of [18] was used with training on the terrestrial and aerial multi-view imagery taking from

30-60 secs per 3D structure extraction (e.g., per plot in the aerial imagery case, per tree in the terrestrial imagery case). Adding the LiDAR constraints was done following the implementation from [21]. The neural radiance architecture is a multilayer perceptron (MLP) with two hidden layers and a ReLU layer per hidden layer and a linear output layer as in [18]. Training was performed using the ADAM optimizer [14] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-15}$ using NVIDIA Tesla V100.

The main limitation of neural fields from aerial multi-view imagery is the presence of occlusion of sub-canopy structure, specially in densely forested areas. In our case, fusion with TLS data can resolve this problem as terrestrial data provides highly detailed sub-canopy information. Additionally, when TLS is unavailable, terrestrial imagery can be used instead. Our 3D structure experiments from terrestrial multi-view information in Sec.3.1 and the DBH estimation performance results demonstrate that highly detailed structure along the entire vertical stand direction can be extracted by neural fields when image information is available. In the absence of multi-view image data, however, neural fields are not capable of generating synthetic information behind occluded areas and performance on metrics affected by occlusion are expected to yield large errors. This problem can be alleviated through multi-view images capturing the desired areas of interest in the ecosystem.

6   Conclusion

In this work, we proposed neural radiance fields as representations that can finely express the 3D structure of forests both in the *in-situ* and at the broad landscape scale. In addition, the properties of neural radiance fields; in particular, the fact that they account for both the origin and direction of radiance to define 3D structure enables the fusion of data coming from multiple locations and modalities; more specifically those from multi-view LiDAR's and cameras. Finally, we evaluated the performance of 3D structure derived metrics typically used in forest monitoring programs and demonstrated the potential of neural fields to improve performance of scalable methods at near the level of *in-situ* methods. This not only represents a benefit on sampling time efficiency but also has powerful implications on reducing monitoring costs.

Acknowledgements

References

1. Atchley, A., Linn, Rodman, J.A., Hoffman, C., Hyman, J.D., Pimont, F., Sieg, C., Middleton, R.S.: Effects of fuel spatial distribution on wildland fire behaviour. International Journal of Wildland Fire **30**(3), 179–189 (2021)

2. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2), 239–256 (Feb 1992). https://doi.org/10.1109/34.121791

3. Castorena, J., Creusere, C.D., Voelz, D.: Modeling lidar scene sparsity using compressive sensing. In: 2010 IEEE International Geoscience and Remote Sensing Symposium. pp. 2186–2189. IEEE (2010)

4. Castorena, J., Dickman, L.T., Killebrew, A.J., Gattiker, J.R., Linn, R., Loudermilk, E.L.: Automated structural-level alignment of multi-view tls and als point clouds in forestry (2023)

5. Castorena, J., Puskorius, G.V., Pandey, G.: Motion guided lidar-camera self-calibration and accelerated depth upsampling for autonomous vehicles. Journal of Intelligent & Robotic Systems **100**(3), 1129–1138 (2020)

6. Dubayah, R.O., Drake, J.B.: Lidar remote sensing for forestry. Journal of forestry **98**(6), 44–46 (2000)

7. FAO, U.: The state of the world's forests 2020. In: Forests, biodiversity and people. p. 214. Rome, Italy (2020). https://doi.org/https://doi.org/10.4060/ca8642en

8. Gao, W., Tedrake, R.: Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11095–11104 (2019)

9. Ge, X., Zhu, Q.: Target-based automated matching of multiple terrestrial laser scans for complex forest scenes. ISPRS Journal of Photogrammetry and Remote Sensing **179**, 1–13 (2021)

10. Hilker, T., van Leeuwen, M., Coops, N.C., Wulder, M.A., Newnham, G.J., Jupp, D.L., Culvenor, D.S.: Comparing canopy metrics derived from terrestrial and airborne laser scanning in a douglas-fir dominated forest stand. Trees **24**(5), 819–832 (2010)

11. Hyyppä, J., Yu, X., Hyyppä, H., Vastaranta, M., Holopainen, M., Kukko, A., Kaartinen, H., Jaakkola, A., Vaaja, M., Koskinen, J., et al.: Advances in forest inventory using airborne laser scanning. Remote sensing **4**(5), 1190–1207 (2012)

12. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH computer graphics **18**(3), 165–174 (1984)

13. Kankare, V., Joensuu, M., Vauhkonen, J., Holopainen, M., Tanhuanpää, T., Vastaranta, M., Hyyppä, J., Hyyppä, H., Alho, P., Rikala, J., et al.: Estimation of the timber quality of scots pine with terrestrial laser scanning. Forests **5**(8), 1879–1895 (2014)

14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

15. Lausch, A., Erasmi, S., King, D.J., Magdon, P., Heurich, M.: Understanding forest health with remote sensing-part ii—a review of approaches and data models. Remote Sensing **9**(2), 129 (2017)

16. Linn, R., Reisner, J., Colman, J.J., Winterkamp, J.: Studying wildfire behavior using firetec. International journal of wildland fire **11(4)**, 233–246. (2002)

17. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. arXiv preprint arXiv:2003.08934 (2020)

18. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127

19. Myronenko, A., Song, X.: Point set registration: Coherent point drift. IEEE transactions on pattern analysis and machine intelligence **32**(12), 2262–2275 (2010)

20. Pokswinski, S., Gallagher, M.R., Skowronski, N.S., Loudermilk, E.L., Hawley, C., Wallace, D., Everland, A., Wallace, J., Hiers, J.K.: A simplified and affordable approach to forest monitoring using single terrestrial laser scans and transect sampling. MethodsX **8**, 101484 (2021)

21. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Niebner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 12892–12901 (2022)
22. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
23. Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E., Gabler, K., Schadauer, K., Vidal, C., Lanz, A., Staahl, G., Cienciala, E.: National forest inventories. Pathways for Common Reporting. European Science Foundation **1**, 541–553 (2010)
24. Vierling, K.T., Vierling, L.A., Gould, W.A., Martinuzzi, S., Clawges, R.M.: Lidar: shedding new light on habitat characterization and modeling. Frontiers in Ecology and the Environment **6**(2), 90–98 (2008)
25. White, J.C., Coops, N.C., Wulder, M.A., Vastaranta, M., Hilker, T., Tompalski, P.: Remote sensing technologies for enhancing forest inventories: A review. Canadian Journal of Remote Sensing **42**(5), 619–641 (2016)
26. Windrim, L., Bryson, M.: Detection, segmentation, and model fitting of individual tree stems from airborne laser scanning of forests using deep learning. Remote Sensing **12**(9), 1469 (2020)
27. Xie, H., Yao, H., Zhou, S., Mao, J., Zhang, S., Sun, W.: Grnet: Gridding residual network for dense point cloud completion. In: ECCV (2020)