Chapter 1
Full-Field 3D Structural Dynamics via Neural Radiance Fields

Thiago F. Ribeiro, Victor H. R. Cardoso, João C. W. A. Costa, and Moisés F. Silva

Abstract The ability to acquire high-resolution, full-field 3D representations of structures has long enabled experimental modal identification, whether via contact sensors that track material points in a Lagrangian frame or optical imagers that sample motion in an approximate Eulerian frame. Conventional algorithms assume fixed measurement identities through time, but large-motion dynamics, occlusions, viewpoint changes, and noise violate both frames and undermine standard techniques. We introduce the first Neural Radiance Field (NeRF)-based pipeline for full-field 3D experimental modal analysis driven solely by standard RGB cameras. A semi-spherical array of synchronized cameras records the structure's oscillation at discrete time steps. At each step, we train a NeRF to reconstruct the instantaneous 3D light field and extract a dense point cloud with a Lagrangian interpretation. The resulting sequence of dynamically varying point clouds is then processed via principal component analysis for dimensionality reduction, followed by a blind source separation scheme to decompose modal frequencies and recover 3D mode shapes. Validation on simulated data yields modal parameters within experimental error margins of reference values, demonstrating accuracy on par with specialized hardware while relying solely on ubiquitous RGB imaging and neural reconstruction. By integrating NeRF reconstructions with classical modal-analysis workflows, this work establishes a cost-effective, vision-based alternative for structural diagnostics. Motivated by the very first attempt to assess the health condition of plants through their vibrational modes, treating them within the framework of structural health monitoring, this work also represents the first application of dynamic Neural Radiance Fields for full-field modal analysis of plants, enabling non-invasive extraction of quantitative physiological traits.

Keywords Neural radiance fields · 3D scanning · BSS · PCA · dynamic cloud points

Introduction

Modal analysis of dynamic structures represents a fundamental methodology in structural engineering and mechanical systems characterization, enabling the identification of natural frequencies, damping ratios, and mode shapes essential for design validation and health monitoring. Traditional contact-based sensor networks impose significant limitations in terms of spatial resolution, installation complexity, and applicability to inaccessible environments. The emergence of vision-based measurement techniques has revolutionized structural dynamics assessment by offering non-contact, high-resolution spatial measurements capable of capturing full-field deformation patterns [1, 2].

Recent advances in computer vision and neural implicit representations have introduced NeRFs as a powerful tool for 3D scene reconstruction and novel view synthesis. NeRFs achieve state-of-the-art results by representing scenes through fully-connected neural networks that map 5D coordinates (spatial location and viewing direction) to volume density and view-dependent radiance [3]. This implicit representation enables high-fidelity 3D reconstruction from sparse 2D observations, making it particularly suitable for capturing dynamic structural behavior with unprecedented spatial detail.

Thiago F. Ribeiro · Victor H. R. Cardoso · João C. W. A. Costa
Applied Electromagnetism Laboratory
Federal University of Pará, Belém, PA, Brazil
e-mail: thiago.figueiro.ribeiro@gmail.com; victorcard@ufpa.br; jweyl@ufpa.br

Moisés F. Silva
Los Alamos National Laboratory
Los Alamos, New Mexico 87544, USA
e-mail: cien.felipe@gmail.com

Kristin B. Zimmerman, PhD (eds.), *Data Science in Engineering, 2026, Vol. 9*, Conference Proceedings of the Society for Experimental Mechanics Series, https://doi.org/10.1007/978-8-743-81448-1_1

The application of 3D point cloud data to structural dynamics has gained significant momentum with the development of time-of-flight (ToF) imaging technologies. The work present in [4] demonstrated the feasibility of extracting vibration modes from dynamic point clouds using commercial depth sensors, employing virtual Lagrangian sensors to track structural motion and blind source separation techniques for modal parameter extraction. This pioneering work established that point cloud-based approaches could achieve high-resolution modal identification comparable to traditional laser displacement sensors, with the advantage of capturing thousands of measurement points simultaneously.

While traditional point cloud acquisition relies on sequential scanning or ToF sensors with limited temporal resolution, NeRF-based reconstruction offers distinct advantages for dynamic structural analysis. Unlike conventional photogrammetry or stereo vision approaches, NeRF technology provides superior handling of complex lighting conditions, view-dependent effects, and scene completeness. The continuous volumetric representation inherent in NeRFs enables the generation of dense, temporally consistent point clouds from multi-view imagery, addressing the fundamental challenge of maintaining spatial correspondence across time-varying datasets.

Principal Component Analysis (PCA) serves as the mathematical foundation for extracting dominant spatial patterns from high-dimensional point cloud datasets. In structural dynamics applications, PCA enables the decomposition of complex vibration patterns into orthogonal principal components corresponding to fundamental vibration modes [5, 6]. For an undamped or lightly damped structure with uniform mass distribution, the principal directions converge to mode shape directions, with singular values indicating modal participation energy [4]. This transformation effectively converts geometric information into a friendly format to classical modal analysis techniques while preserving spatial richness.

Blind Source Separation (BSS) complements PCA by addressing the fundamental challenge of extracting independent modal responses from mixed observations without prior knowledge of source characteristics. BSS-based methods for modal identification treat modal coordinates as independent sources, enabling separation of closely-spaced modes and identification of operational deflection shapes under ambient excitation [7, 8]. The complexity pursuit algorithm has demonstrated particular effectiveness for estimating closely spaced and highly damped modes with minimal user supervision [9], making it suitable for processing high-dimensional point cloud data.

Current research gaps exist at the intersection of neural implicit representations and structural dynamics characterization. While NeRF technology has demonstrated success in static scene reconstruction, its application to dynamic structural systems remains largely unexplored. The temporal consistency requirements for modal analysis demand robust tracking of geometric features across multiple time instances, which challenges traditional NeRF formulations optimized for static scenes. Additionally, the applications of NeRF-based point cloud sequences for real-time structural monitoring requires systematic investigation.

The integration of NeRF-based 3D reconstruction with advanced signal processing techniques represents a significant advancement beyond existing point cloud-based modal analysis methods. Traditional approaches using ToF sensors are limited by frame rate constraints and depth resolution, restricting applications to structures with relatively low natural frequencies [4]. NeRF-based approaches, leveraging multi-view imagery with higher temporal resolution of high-speed cameras, could potentially overcome these limitations while providing enhanced spatial coverage and measurement density.

This paper presents a novel methodology that leverages NeRF-based 3D reconstruction for generating high-resolution point cloud representations of vibrating structures, followed by PCA-BSS processing for modal parameter extraction. The approach addresses fundamental limitations in current experimental modal analysis practices by combining the spatial richness of neural implicit representations with established blind identification techniques. The methodology extends recent advances in dynamic point cloud processing [4] by introducing NeRF-based reconstruction as a means to achieve superior temporal consistency and spatial resolution.

The proposed approach represents a paradigm shift from sparse sensor networks to dense, spatially distributed measurement systems capable of capturing intricate mode shape details that conventional accelerometer arrays cannot resolve. By integrating cutting-edge computer vision techniques with established structural dynamics theory, this work opens new frontiers in non-contact structural health monitoring with implications extending to robotics, aerospace, and biomedical applications.

While Neural Radiance Fields (NeRF) [3] have been successfully applied to the static 3D reconstruction of plants [10] and civil structures [11], their extension to dynamic scenes for quantitative modal analysis remains unexplored. This work presents, to the best of our knowledge, the first methodology to leverage the dense, view-consistent spatiotemporal reconstructions from dynamic NeRF for full-field operational modal analysis.

The application of Neural Radiance Fields (NeRF) to represent the spatiotemporal dynamics of plant vibration modes presents a novel methodology for the non-invasive extraction of quantitative physiological traits. This technique yields high-resolution, full-field quantitative data *in planta*, establishing a foundation for scalable phenotyping at the canopy level. In contrast to established but often cost-prohibitive and range-limited LiDAR systems, the proposed method leverages standard high-definition camera systems, enabling unprecedented data fidelity at a substantially reduced operational cost.

Our primary contributions include: (1) demonstration of NeRF applicability to dynamic structural systems for modal analysis applications, (2) development of processing workflows that balance computational efficiency with measurement fidelity, and (3) validation of the methodology against simulation based ground truth data. The results provide a foundation for advancing vision-based structural dynamics assessment and highlight the transformative potential of neural implicit representations in experimental mechanics.

Neural Radiance Fields

Neural radiance fields represent scenes as continuous volumetric functions, learned through differentiable rendering of multi-view images [12]. Unlike traditional photogrammetric approaches based on Structure-from-Motion (SfM) and Multi-View Stereo (MVS), NeRF-based methods demonstrate superior performance in handling complex geometries, reflective surfaces, and sparse viewpoints [13]. The extraction of dense point clouds from trained NeRF models provides geometrically consistent three-dimensional representations suitable for downstream structural analysis tasks [14].

MLP architecture

The NeRF architecture employs a multilayer perceptron (MLP) consisting of nine fully connected layers, each with 256 neurons and ReLU activation functions. A skip connection is incorporated into the fifth layer, where positional encoding [15] is introduced to enhance the network's ability to capture fine spatial details, as depicted in Figure 1. The ninth layer outputs the density parameter, σ. An additional tenth layer, with 128 neurons and ReLU activation, processes the remaining 256-dimensional feature vector. A final sigmoid-activated output layer predicts the emitted RGB radiance at position x, as observed along a ray with direction d [3].

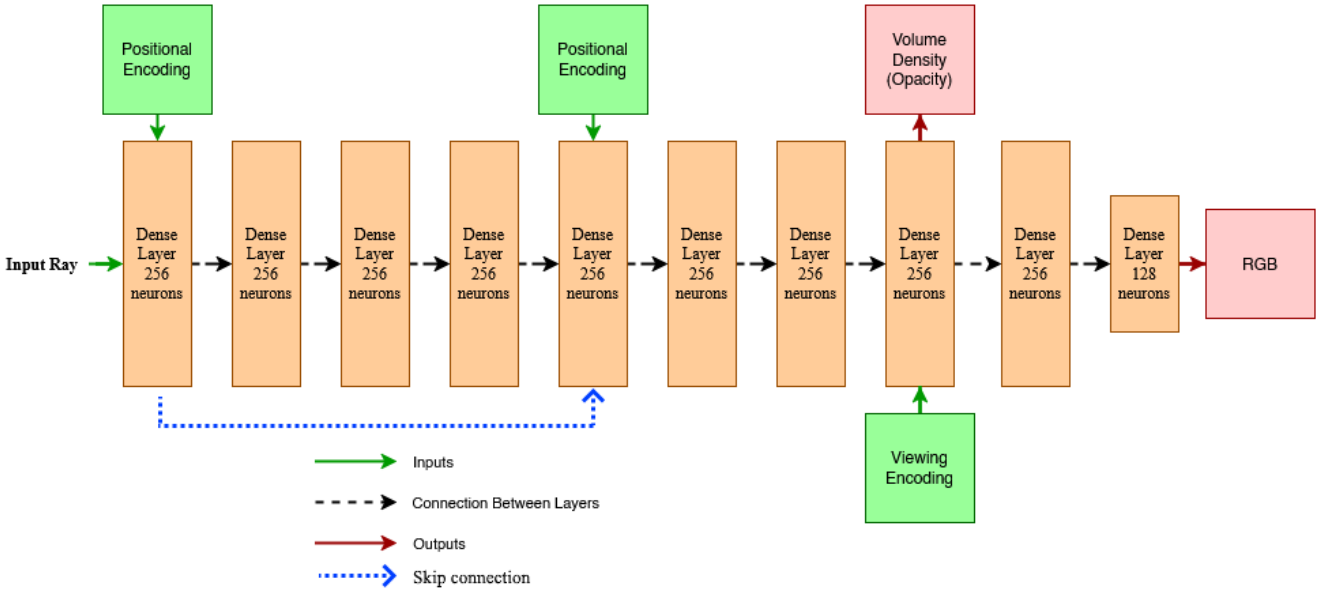

Fig. 1 Visualization of the NeRF's fully-connected network architecture.

Volume rendering

The author in [3] models the 5D neural radiance field to represent a scene as the volume density and directional emitted radiance at any point in space. The color of any ray passing through the scene is rendered using principles from classical volume rendering [16]. The volume density $\sigma(\mathbf{x})$ is interpreted as the differential probability of a ray terminating at an infinitesimal particle at location $\mathbf{x}$. Equation 1 depicts the expected color $C(\mathbf{r})$ of camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with near and

far bounds t_n and t_f is:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))c(\mathbf{r}(t), d)dt, \quad T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right). \tag{1}$$

The function $T(t)$ denotes the accumulated transmittance along the ray from t_n to t, i.e., the probability that the ray travels from t_n to t without hitting any other particle. Rendering a view from a continuous neural radiance field requires estimating this integral $C(\mathbf{r})$ for a camera ray traced through each pixel of the desired virtual camera.

This continuous integral can be numerically estimated using quadrature. Deterministic quadrature, which is typically used for rendering discretized voxel grids, would effectively limit the representation's resolution because the MLP would only be queried at a fixed discrete set of locations. Instead, a stratified sampling approach is used where its partitioned $[t_n, t_f]$ into N evenly-spaced bins and then draw one sample uniformly at random from within each bin, as depicted in Equation 2:

$$t_i \sim \mathcal{U}\left[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)\right]. \tag{2}$$

Although a discrete set of samples is used to estimate the integral, stratified sampling enables representing a continuous scene representation because it results in the MLP being evaluated at continuous positions over the course of optimization. These samples are used to estimate $C(\mathbf{r})$ with the quadrature rule discussed in the volume rendering review by Max [17]:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))c_i, \quad T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j\delta_j\right), \tag{3}$$

where $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples. This function for calculating $\hat{C}(\mathbf{r})$ from the set of (c_i, σ_i) values is trivially differentiable and reduces to traditional alpha compositing with alpha values $\alpha_i = 1 - \exp(-\sigma_i\delta_i)$.

Positional encoding

Positional encoding[1] is a method used to inject spatial information into neural networks, enabling them to effectively model high-frequency variations and complex spatial structures. It maps low-dimensional inputs, such as spatial coordinates (x, y, z) or viewing directions (θ, ϕ), into a higher-dimensional feature space using a series of sinusoidal functions. This technique was introduced in the context of Transformers [15] and has found applications in Large Language Models [18], audio processing [19] and graph-structured data [20]. It was also adapted for use in NeRFs [3], where capturing fine-grained spatial details is essential.

The positional encoding function is defined in Equation 4:

$$\text{PE}(p) = \left[\sin(2^k\pi p), \cos(2^k\pi p)\right]_{k=0}^{L-1} \tag{4}$$

where p represents the input coordinate (e.g., x, y, z), and L determines the number of frequency bands used in the encoding [15]. The canonical NeRF uses $L = 10$ frequencies for spatial coordinates (x, y, z) and $L = 4$ for viewing directions (θ, ϕ).

For a single input p, the positional encoding expands it into a $2L$-dimensional feature vector:

$$\text{PE}(p) \in R^{2L}. \tag{5}$$

For 3D spatial coordinates (x, y, z), the encoded representation is computed as:

$$\text{PE}(x, y, z) = [\text{PE}(x), \text{PE}(y), \text{PE}(z)], \tag{6}$$

resulting in a $6L$-dimensional feature vector for each 3D point. Similarly, for viewing directions (θ, ϕ) the encoded representation is:

$$\text{PE}(\theta, \phi) = [\text{PE}(\theta), \text{PE}(\phi)], \tag{7}$$

Positional encoding is crucial in overcoming the limitations of standard neural networks, which struggle to model high-frequency signals due to the inherent smoothness of activation functions like ReLU. By incorporating Fourier features, the network gains the capacity to represent fine details and sharp edges in data, significantly improving the quality of reconstructions and representations in tasks like 3D rendering and image synthesis.

[1] Positional encoding was inspired by Fourier feature mapping [3], a technique used to enable neural networks, especially MLPs, to better capture high-frequency signals in data, particularly in low-dimensional domains such as image and graphics processing.

Hierarchical volume sampling

NeRF's rendering strategy of densely evaluating the neural radiance field network at N query points along each camera ray is inefficient: free space and occluded regions that do not contribute to the rendered image are still sampled repeatedly. An hierarchical representation is proposed by [3] that increases rendering efficiency by allocating samples proportionally to their expected effect on the final rendering.

Instead of just using a single network to represent the scene, two networks are simultaneously optimized: one "coarse" ($N_c = 64$) and one "fine" ($N_f = 192$). First a set of N_c locations are sampled using stratified sampling, and evaluate the "coarse" network at these locations as described in equations (2) and (3). Given the output of this "coarse" network, it is then produced a more informed sampling of points along each ray where samples are biased towards the relevant parts of the volume, as depicted in Figure 2. To do this, first the alpha composited color is rewritten from the coarse network $\hat{C}_c(\mathbf{r})$ in equation (3) as a weighted sum of all sampled colors c_i along the ray, depicted in Equation 8:

$$\hat{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} w_i c_i, \qquad w_i = T_i(1 - \exp(-\sigma_i \delta_i)). \tag{8}$$

Normalizing these weights as $\hat{w}_i = w_i / \sum_{j=1}^{N_c} w_j$ produces a piecewise-constant Probability Density Function (PDF) along the ray. A second set of N_f locations is sampled from this distribution using inverse transform sampling, evaluate the "fine" network at the union of the first and second set of samples, and compute the final rendered color of the ray $\hat{C}_f(\mathbf{r})$ using the equation (3) but using all $N_c + N_f$ samples. This procedure allocates more samples to regions expected to contain visible content. This addresses a similar goal as importance sampling, but the sampled values are used as a non-uniform discretization of the whole integration domain rather than treating each sample as an independent probabilistic estimate of the entire integral.

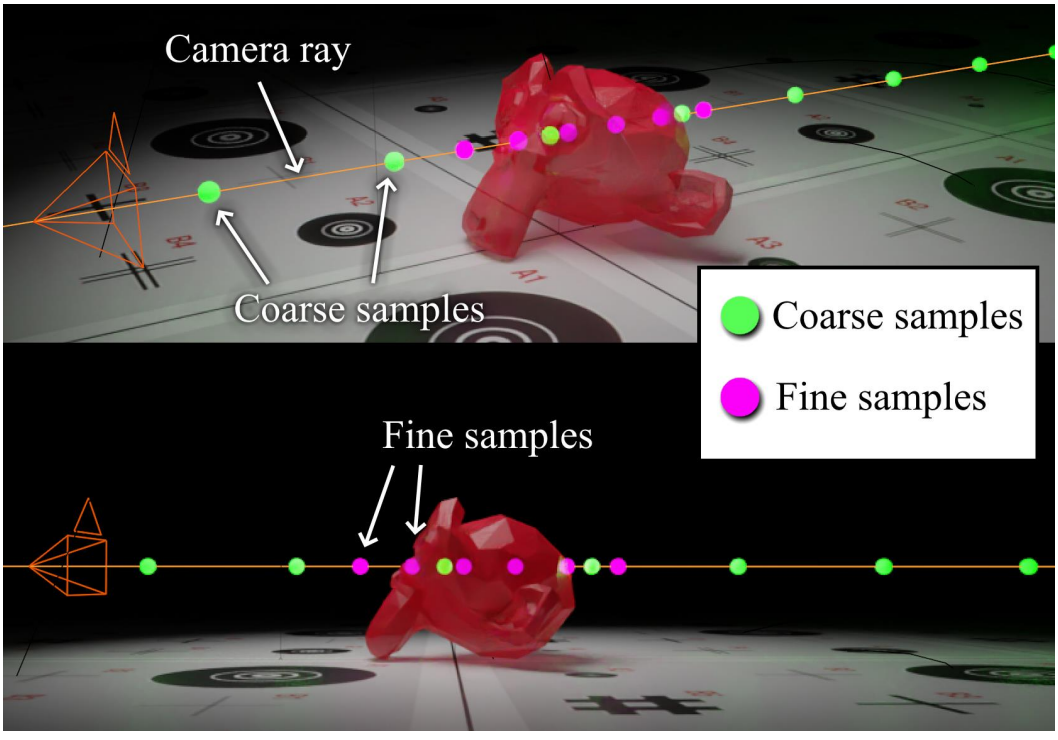

Fig. 2 Visual representation of NeRF's hierarchical sampling. First a set of coarse points (in green) are uniformly sampled along a particular ray using a coarse network to estimate general scene density. This initial density prediction is then used to form a probability distribution that guides the fine network to finely sample additional points (in pink) in regions that likely contain important details.

Loss function

During training of deep learning models, a loss function is used to optimize the model's parameters [21]. It measures the difference between the predicted and expected outputs of the model. The objective of training is to minimize this difference as the training iterations occur. Given the vast array of deep learning applications, different loss functions can be applied to different kind of problems. For instance, using 0-1 loss, perceptron loss, logarithmic loss, sigmoid cross entropy loss, and others for classification problems and Absolute loss, Huber Loss, Log-cosh loss and others in regression problems [22]. In image models, the loss function is computed by comparing the colors of a synthesized image with the ground-truth image [23]. Traditional image models rely on the L2 norm, also known as the Euclidean Norm, for image reconstruction [24].

In the context of NeRF training, the Loss function defined at Equation 9 was used to measure the difference between a predicted image $\hat{C}_p(r)$ and a ground truth image $C(r)$. The loss function of the NeRF is the total squared error between the rendered and true pixel colors for both the coarse $(N_c = 64)$ and fine $(N_f = 192)$ renderings:

$$L = \sum_{r \in R} \left[\|\hat{C}_c(r) - C(r)\|_2^2 + \|\hat{C}_f(r) - C(r)\|_2^2 \right]. \tag{9}$$

NeRFs are trained on batches of rays rather than entire images. R represents the set of rays in each training batch. For each ray r in a batch, $C(r)$ denotes the ground truth RGB color, while $\hat{C}_c(r)$ and $\hat{C}_f(r)$ represent the predicted RGB colors from the coarse and fine models, respectively. The training aims minimizing the difference between these predicted colors and the ground truth, ensuring that both coarse and fine models learn to accurately reconstruct the scene.

Point cloud extraction and export

Extracting explicit geometry information in the shape of a point cloud from an implicit volumetric representation of a scene from a trained NeRF model is done by sampling a batch of rays from different viewing angles. Using these rays, the NeRF model infers their RGB color and opacity for every sampled 3D coordinate point. Since the camera origins and directions of every ray are known, the distance/depth of every sampled point to the camera origin are also known from the camera extrinsics. The distance/depth from the camera's origin to the sample point is encoded into 3D coordinates in space.

$$P = \{x_i \mid \sigma(x_i) > \sigma_{threshold} \text{ for } x_i \in \mathbb{R}^3\}, \tag{10}$$

where $\sigma(x_i)$ is the predicted density at point (x_i) and $\sigma_{threshold}$ is a predefined opacity threshold used to extract surface points. Points below certain $\sigma_{threshold}$ (often 0.9) or outside a bounding box are discarded. The remaining valid points are gradually collected, with the process repeating until an arbitrary number of high quality points have been gathered. Finally, the accumulated 3D points and their color data can be exported into a standardized point cloud format for further processing. An advantage of implicit representations is the continuous nature of the reconstructed model. Therefore, given sufficient memory resources, point clouds containing an arbitrary number of points can be generated from any region of the model.

Motion Tracking by Virtual Lagrangian Sensors

Given a sequence of time-varying point cloud data, we define a densely populated grid of three-dimensional points (Figure 3). Each point in this grid is used as a measurement point for the motion tracking procedure. The goal is to generate virtual Lagrangian sensors by, allowing the efficient measurement of displacements at thousands of points across the structure's surface. In the case of Figure 3, we are able to densely monitor the structure using 99.990 virtual Lagrangian sensors for dynamic point cloud data with approximately 100.000 points. Note that the number of points in the point cloud data is not constant across time. As a result this modelling step is imperative to enable the measurement of motion in the same structure locations over time. Also, the number of virtual sensors can be arbitrarily chosen, and randomly sampled from the model. The concern is to sample the same points across time to simulate real sensors placed on the structure's surface as the points in the dynamic cloud are not suitable to directly use since the number of points change dynamically.

The dense number of sensing points is achieved by fitting a three-dimensional polynomial model over the point cloud data. A fixed set of points are sampled from the X- and Y-directions to measure displacements in the Z direction over time. Thus, motion is tracked on the same locations of the structure once the values for the Z-axis are interpolated by fitting the model. This process is repeated across time by fitting the three-dimensional model for each of the acquired point cloud data. Note that the model does not use any information regarding structure's properties or dimensions, instead, this unsupervised model is purely data-driven.

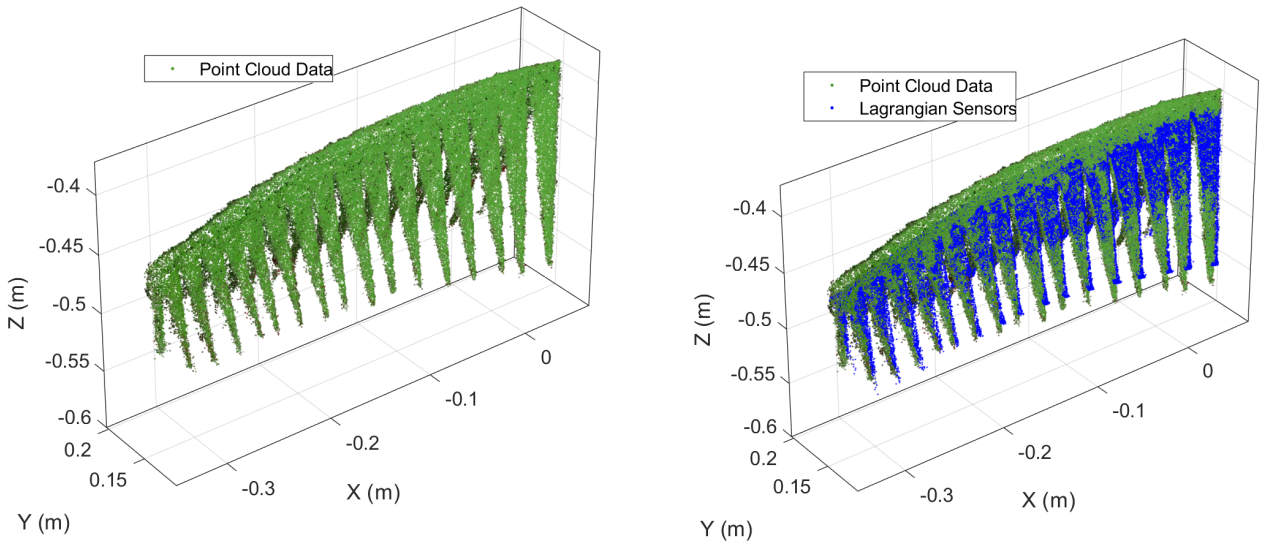

Fig. 3 Dynamic point cloud of a vibrating *Euterpe oleracea Mart.* (açaí palm) tree leaf (left) and a very dense grid of Lagrangian sensors on top of the point cloud data (right). In this case, we created 99.990 sensing points, which allows for full-field 3D monitoring for the entire leaf.

The displacement time series are the difference between the square root of the current position, z^i_{new} of the ith virtual sensor to its previous one, z^i_{old}, as in Equation (11), corresponding to a Lagrangian-type measurement once the values of the structure coordinates do not vary along time. Therefore, the end result of the motion tracking is a matrix $\Delta_z \in \mathbb{R}^{n\times m}$ with displacement measurements, where n is the number of dynamic point clouds (temporal dimension) and m the number of virtual Lagrangian sensors (spatial dimension).

$$\Delta^i_z = z^i_{new} - z^i_{old}. \tag{11}$$

Blind Identification of 3D Vibration Modes

Dimensionality Reduction

Once the vibration motion Δ_z is encoded into the motion time series from the dynamic point cloud data, we can directly apply conventional modal identification techniques over the collected time series. Here, we employ a recent developed class of algorithms for blind identification and visualization of vibration modes from video measurements [25, 26, 27].

Since our ultimate goal is to perform high resolution modal identification, it is desirable to use hundreds of virtual Lagrangian sensors. Thus, the first step for modal identification is to perform dimensionality reduction on the basis of Principal Component Analysis (PCA) over the motion time series. The goal is to factorize Δ_z in terms of two unknown matrices $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ such that $\mathbf{U}$ and $\mathbf{V}$ are the left- and right-singular vectors obtained by the eigenvalue decomposition (EVD) of the covariance matrix from Δ_z through Singular Value Decomposition (SVD) [28].

From a structural dynamics point of view [25, 26], it has been shown that for an undamped or very lightly damped structure, if its mass matrix is proportional to the identity matrix (i.e., uniform mass distribution), then the principal directions will converge to the mode shape direction [29, 27] with the corresponding singular values indicating their participating energy in the structural vibration responses, which means that the active structural modes are projected onto a very small number of r principal components

$$\eta \approx \Delta_z \mathbf{U}^T_r, \tag{12}$$

where $\mathbf{U}_r$ is the first $r(\ll n)$ columns of $\mathbf{U}$, and T denotes the transpose operation.

At this point, it is important to highlight the role being played by both $\mathbf{U}_r$ and η. As PCA is applied over the motion time series, $\eta \in \mathbb{R}^{n\times r}$ accounts for a temporal decomposition with each column η_i indicating the time pattern of a specific vibration mode. On the other hand, $\mathbf{U}_r$ incorporates a spatial decomposition with each row $\mathbf{U}_i$, when properly reshaped, resembling the corresponding vibration mode shape. Thus, when considering both η and $\mathbf{U}_r$ we have a 2D spatio-temporal decomposition.

Blind Identification of 3D Modal Parameters

Although η accounts for the vibration patterns extracted from the dynamic point cloud data, as mentioned in [29] and also highlighted by [27], the uniform mass distribution assumption is usually not satisfied, which makes each component η_i still a mixture of modes and can be approximated as a linear combination of the modal coordinates

$$\eta = \mathbf{q}\Psi = \sum_{i=1}^{r} q_i \psi_i, \tag{13}$$

where $\mathbf{q} \in \mathbb{R}^{n \times r}$ is the modal response matrix with q_i being the i-th modal coordinate, and $\Psi^{r \times r}$ is the mixing matrix. Using the relationship between the extracted components and the modal coordinates observed in [27] leads to

$$\Delta_z \approx \mathbf{q}\Psi\mathbf{U}_r = \mathbf{U}_r \sum_{i=1}^{r} q_i \psi_i = \sum_{i=1}^{r} q_i \langle \psi_i \mathbf{U}_r \rangle. \tag{14}$$

Comparing Equation (14) to Equation (13), high resolution mode shapes with resolution equal to the number of virtual sensors are given by

$$\Phi = \mathbf{U}_r \Psi^{-1}, \tag{15}$$

where $\Phi \in \mathbb{R}^{r \times m}$ is the mode shape matrix with each ϕ_i as the i-th mode shape. Similarly, modal coordinates can be estimated by

$$\mathbf{q} = \mathbf{U}_r \Psi. \tag{16}$$

The mixing matrix Ψ can be estimated by simply applying a BSS technique to solve the linear mixture problem stated in Equation (13). As previously demonstrated [25, 26], the complexity pursuit (CP) algorithm [30] can very efficiently estimate closely spaced and highly damped modes with small user supervision, and is therefore adopted here. Other modal parameters, such as resonant frequencies and damping ratios, can be further estimated using conventional Fourier transform and logarithmic decrement techniques.

Note that in this case the mode shapes in Φ are related to the motion in the Z-axis. Thus, in order to visualize a three-dimensional mode shape, the fixed X- and Y-coordinates for the virtual sensors can be used along with the mode shapes in the Z-axis for visualization of 3D modes.

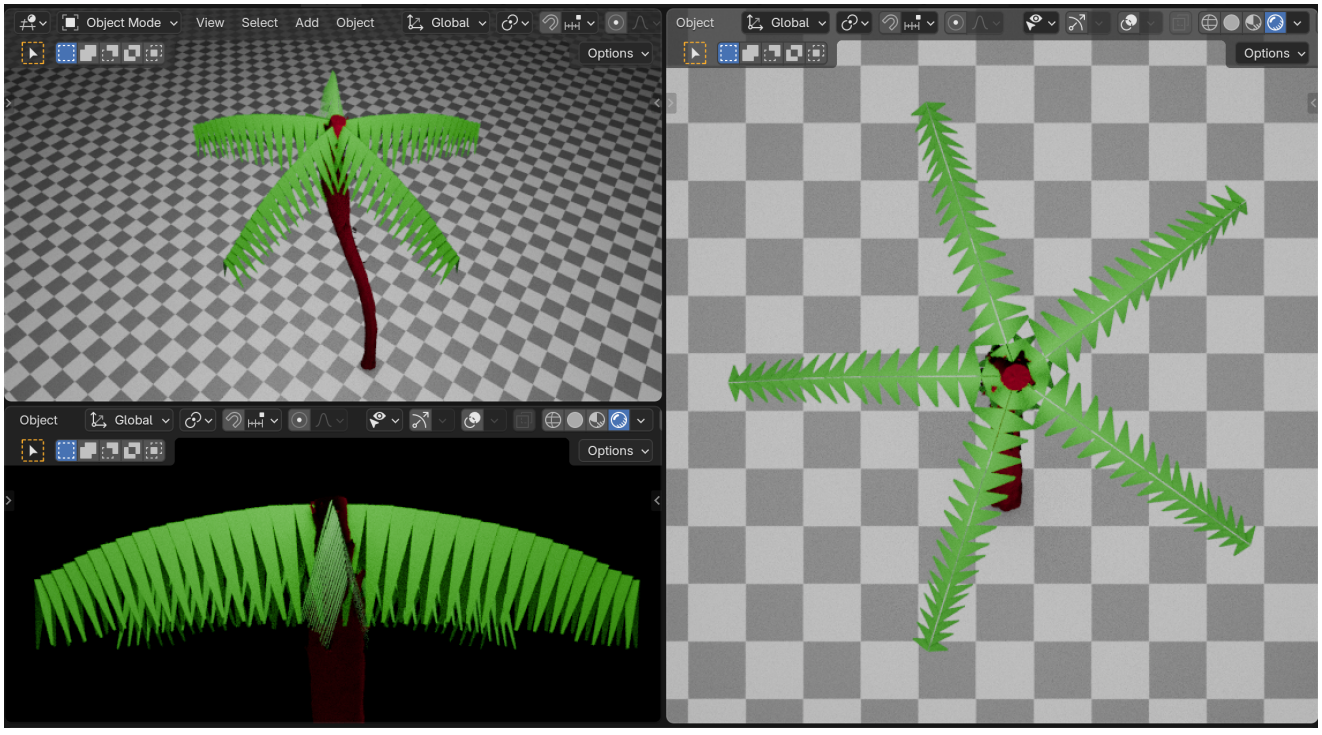

Fig. 4 Three-dimentional model of an *Euterpe oleracea Mart.* (açaí palm), modeled in Blender.

Modal Dynamics Visualization

To reconstruct each independent mode for visualization only, we can use Equation (13) and Equation (12) for a given mode i. First, multiplying the individual modal coordinate q_i in Equation (13) by a magnification factor α and the other ones by a decaying factor β we have

$$\hat{\eta}_i = (\alpha q_i)\psi_i + \beta \sum_{\substack{j=1 \\ j \neq i}}^{r} q_j \psi_j, \tag{17}$$

where $\hat{\eta}_i$ is the reconstructed temporal component after reconstruct the i-th mode. After that we can utilize the mode shape matrix Φ as our new set of spatial components, substitute into Equation (12) and then reconstruct a scaled version of the original measurement with the i-th mode enhanced by applying

$$\hat{\Delta}_z = \hat{\eta}_i \Phi. \tag{18}$$

The end result is a scaled version of the original video by a factor α while removing the other modes using a decaying term β. Other modes can be arbitrarily recovered for visualization purposes following the same procedure.

Estimating 3D Modal Parameters from NeRF-Based Dynamic Point Cloud Data

Data simulation with Blender

A simulation test was conducted to estimate modal parameters from a 3D *Euterpe oleracea Mart.* (açaí palm) tree measuring 2.59 meters long and 1.61 meters wide. The 2.59 meters long trunk geometry was generated using a cylindrical mesh with applied curvature modifiers, while the 0.98 meters long pinnate leaves were constructed through instanced meshes with simple extrusion and array operations to reproduce leaflet distribution. The scene is rendered with a basic PBR shading setup under HDRI illumination on a checkered reference plane, allowing evaluation of form, scale, and topology from multiple perspectives.

An açaí tree with 5 leaves and a branch was modeled. The leaves were modeled simulating z-axis vibration modes of 2hz (first bending mode), 5hz (second bending mode) and 10 hz (first torsional mode) while the branch had a single x-axis 1hz bending mode. 50 pictures from specific points were captured at each moment between 0 and 3 seconds (30 FPS). At each timestep, those 50 Full HD (1920p) pictures were used to train a nerf model which was then used to export an point cloud of that particular moment, totalizing 90 point clouds. Those point clouds were finally used as input for the PCA and BBS process in order to extract its vibration modes.

Similarly to [4] work, 3 principal components were extracted to estimate the first three structural modes. Figure 5 shows that there are 5 active eigenvalues. The presence of additional eigenvalues can be attributed to several inherent properties of the NeRF reconstruction process, such as:

1. Volumetric sampling artifacts arising from the discrete ray-casting procedure;
2. Temporal inconsistencies in point cloud density across timesteps, and/or;
3. Geometric reconstruction noise particularly pronounced at leaf boundaries and complex surface features.

Rather than representing spurious modes, these additional components capture the stochastic nature of the reconstruction process and provide a quantitative measure of method uncertainty.

Instead of fully training nerf models for each 90 timesteps 20k iterations, a transfer learning approach was done. The nerf model at timestep t=0 was initially trained for 17k iterations then for 3k more. The neural network weights for the initial 17k were used as starting weights for subsequent timestep models, each training an additional 3k iterations.

For the leaves, a 5x5 poly function was used to fit the points into an polinomial surface. For the branch, a cylinder fit was used.

Experimental results and discussion

Figure 6 and 7 shows the vibration modes extracted from the nerf-based point cloud models with PCA and BSS respectively. In this case, the resonant frequency estimated for the first bending mode was correctly identified at 2 Hz, while the second bending mode and the first torsional mode frequencies were identified as 5 Hz and 10 Hz.

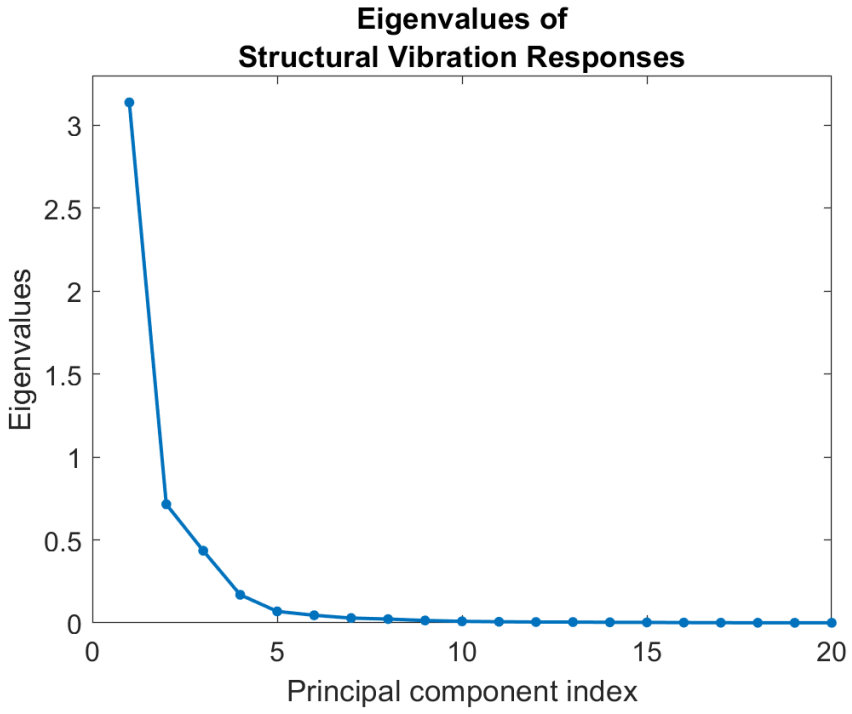

Fig. 5 The eigenvalue distribution of the covariance matrix of the structural responses from the proposed vision approach.

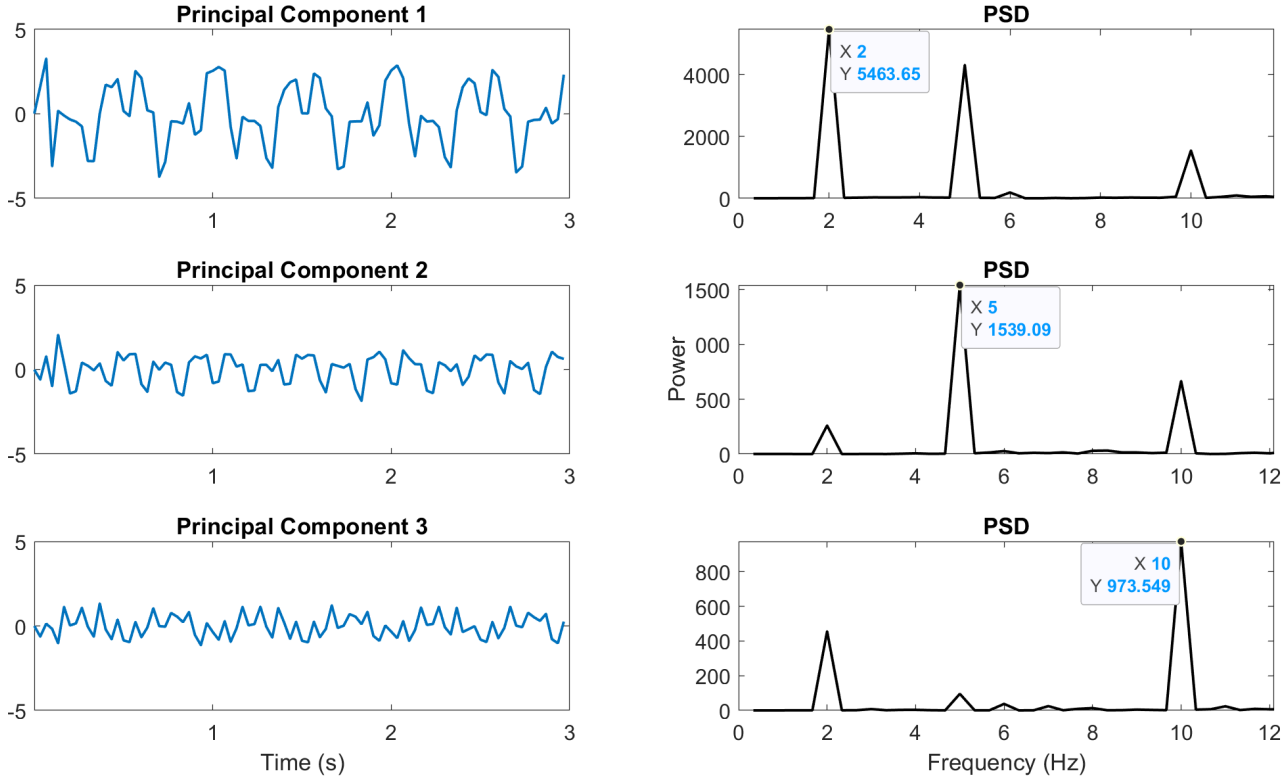

Fig. 6 The principal components and their power spectral density (PSD) of the structural responses from the proposed NeRF-based approach.

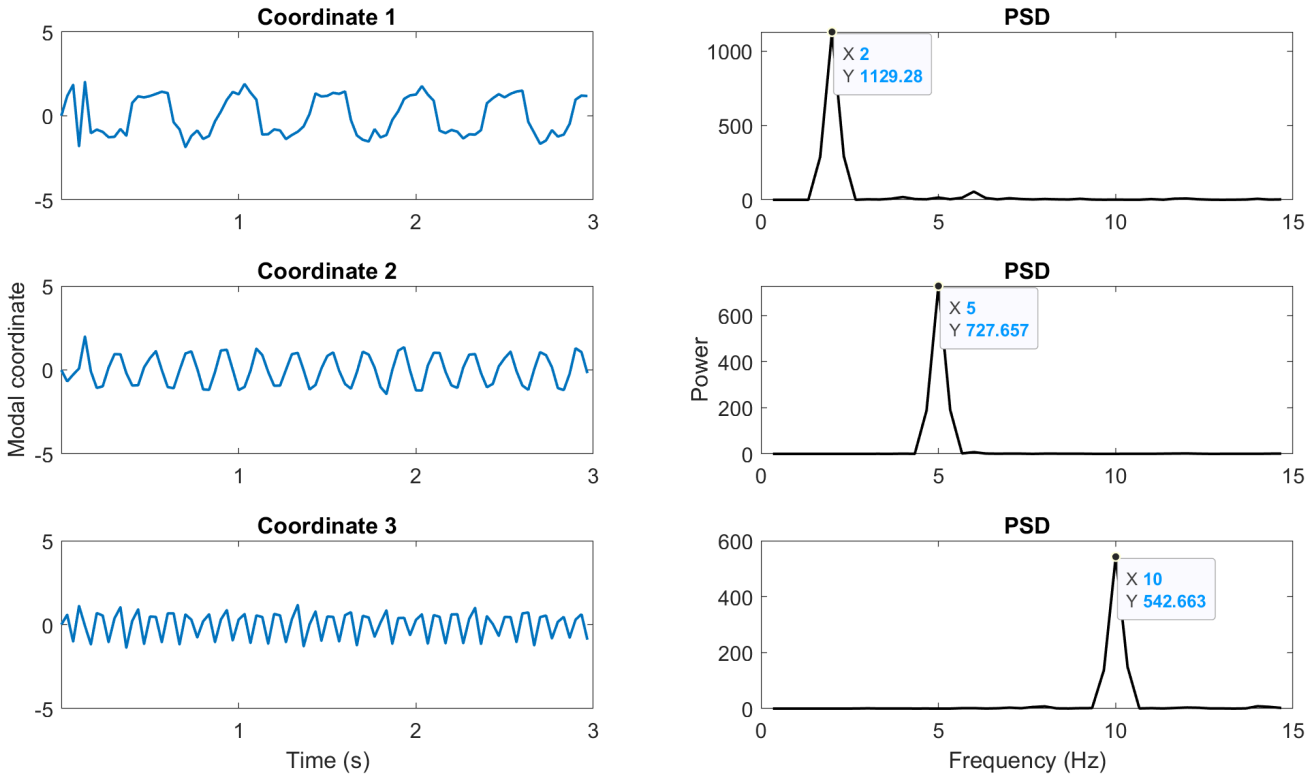

Fig. 7 Modal coordinates and their power spectral densities (PSD) estimated using the blind identification approach from the NeRF-based approach.

The mode static mode shapes are shown in Figure 8. For illustration purposes, several frames from the independent 3D mode shapes can be observed in Figure 9. For comparison purposes, the amplitude of the mode shapes were normalized to the same range. Note that those mode shapes are in perfect agreement with the expected simulated geometry and behavior for their respective modes.

Based on the experimental results, the proposed NeRF-based approach (as depicted in Figure 10) successfully extracted modal parameters from the simulated *Euterpe oleracea Mart.* palm structure. The method correctly identified all three target frequencies: the first bending mode at 2 Hz, second bending mode at 5 Hz, and first torsional mode at 10 Hz, demonstrating accurate frequency domain performance. Both PCA and BSS techniques yielded consistent results in extracting the three principal vibration modes from the 90 reconstructed point clouds. The recovered 3D mode shapes exhibited excellent agreement with the expected simulated geometry, validating the geometric accuracy of the approach. The presence of five active eigenvalues, rather than the expected three, was appropriately attributed to inherent reconstruction artifacts including volumetric sampling noise, temporal density variations, and geometric reconstruction uncertainties at complex features such as leaf boundaries.

The transfer learning strategy employed for NeRF training represents a computationally efficient approach, initializing subsequent timestep models with 17k pre-trained iterations and requiring only 3k additional iterations per frame. This methodology significantly reduces computational burden while maintaining reconstruction quality across the temporal sequence. The integration of polynomial surface fitting (5x5) for leaf structures demonstrates appropriate geometric parameterization for the distinct morphological components. The successful extraction of modal parameters from synthetically generated data establishes a proof-of-concept for vision-based structural health monitoring of vegetation, though real-world validation will require addressing additional challenges including environmental noise, varying illumination conditions, and the stochastic nature of natural wind excitation.

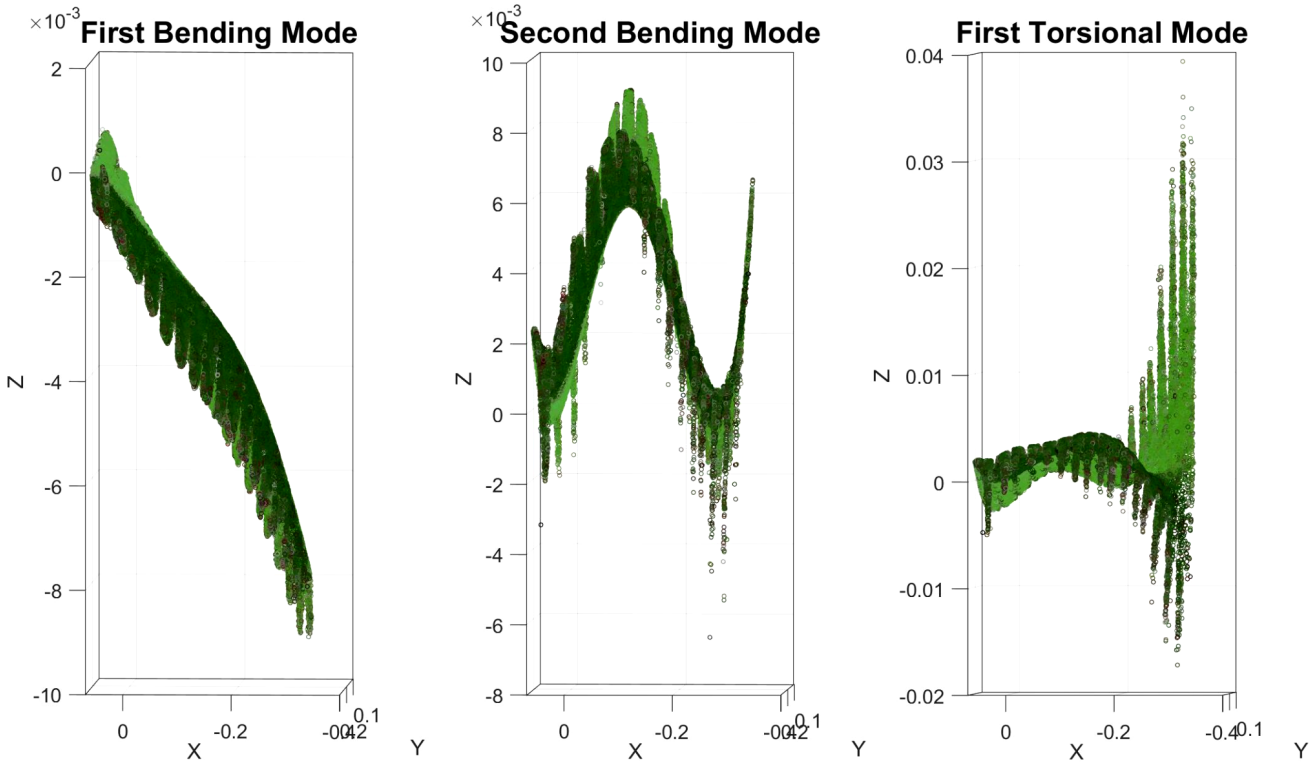

Fig. 8 3D mode shapes estimated from the point cloud data using the blind identification approach based on PCA and blind source separation.

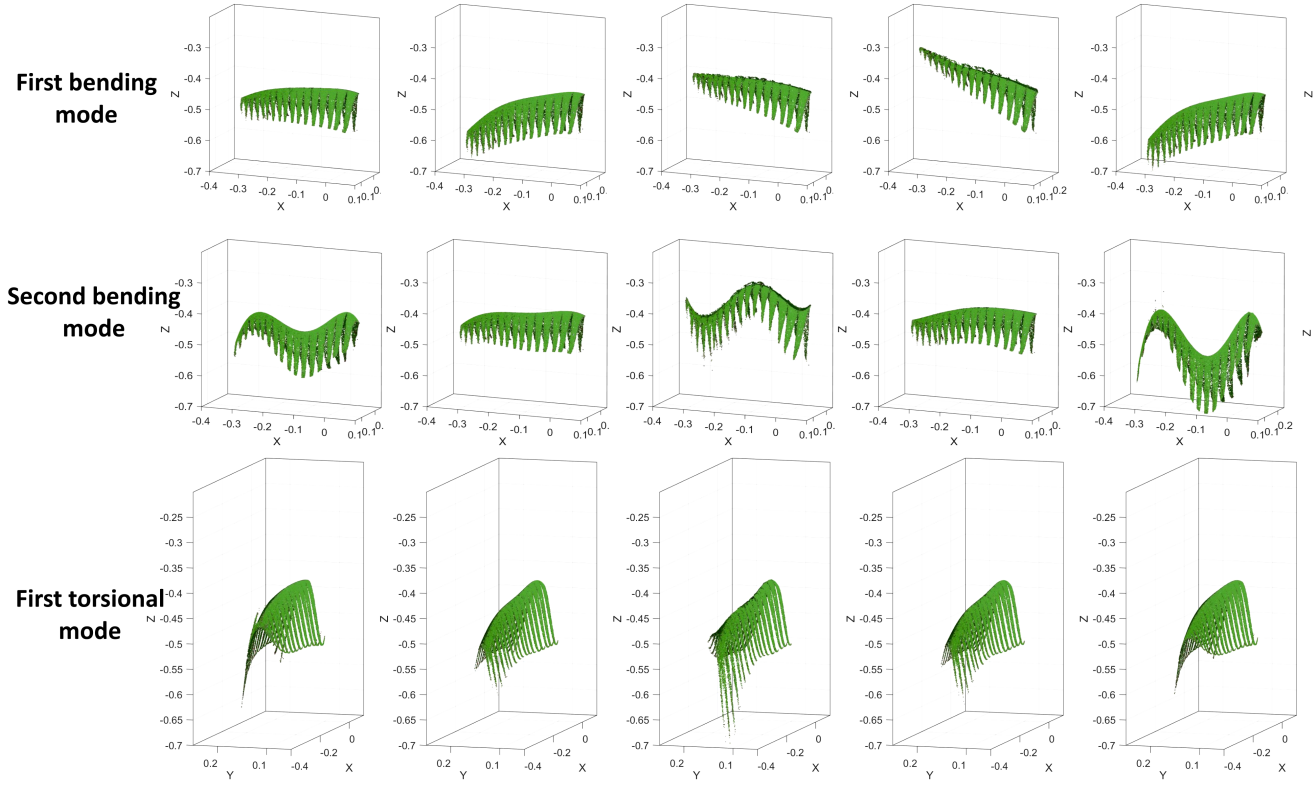

Fig. 9 A few temporal frames from the motion magnified 3D videos for each one of the vibration modes.

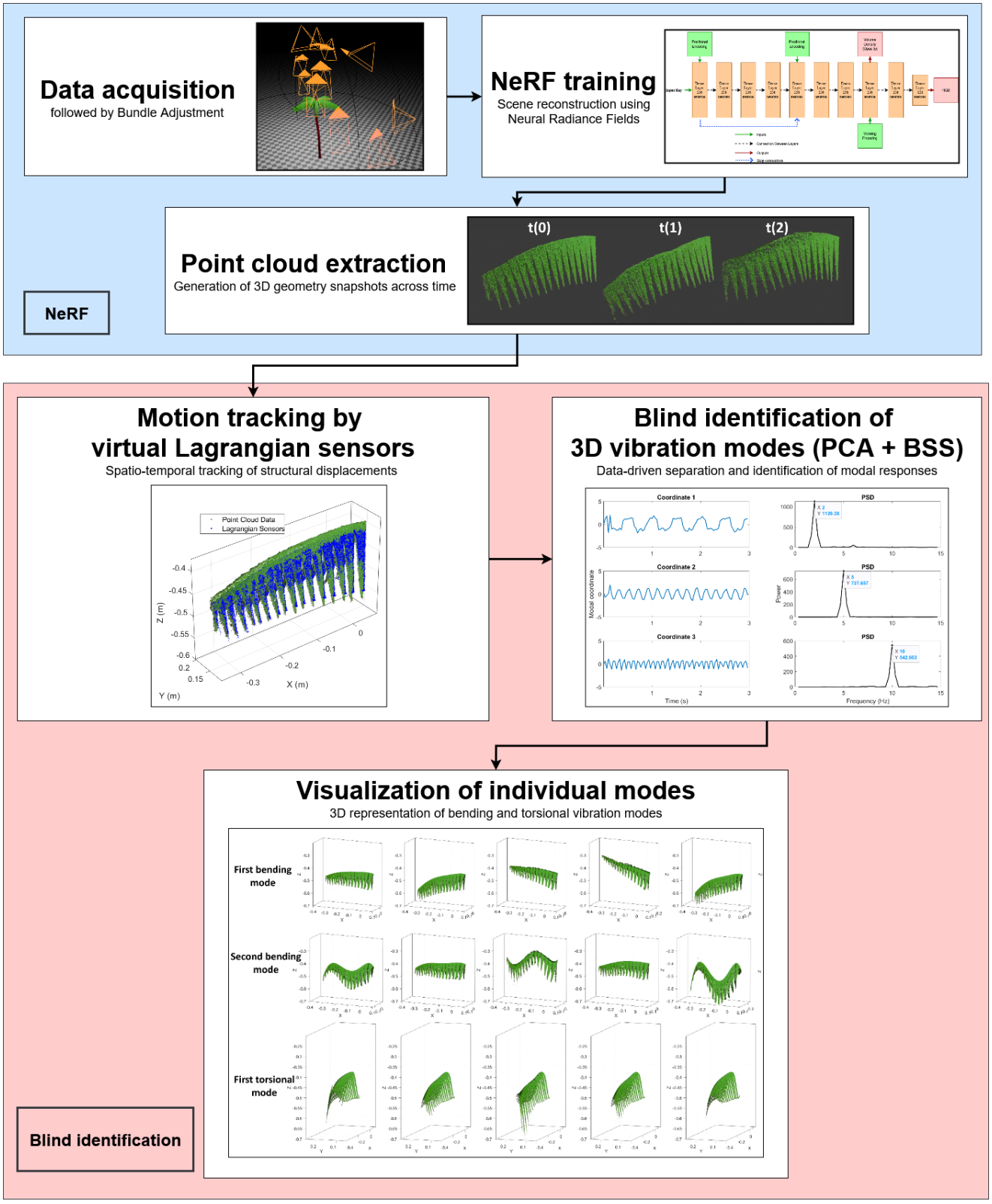

Fig. 10 We introduce the first attempt to process and extract 3D modal parameters from dynamic, NeRF-based, point cloud data. Initially, a set of pictures are obtained from a timestep of the structure. Then, these pictures are input into the NeRF pipeline where, after the training is finished, a point cloud of that timestep is obtained. This process is repeated until all models from the timeframe are trained and their point clouds extracted. Afterwards, blind identification of vibration modes is performed using PCA and BSS algorithms over motion time series estimated for each one of the coordinates of the structure. Finally, the mode shapes and modal coordinates are used to reconstruct the motion of the independent vibration modes.

Conclusion

The integration of machine learning-based 3D reconstruction with established modal analysis techniques represents a paradigm shift in structural dynamics characterization. This convergence enables unprecedented spatial and temporal resolution while maintaining the theoretical foundations of modal decomposition. The proposed NeRF-PCA-BSS pipeline establishes a foundation for next-generation structural health monitoring systems, combining the geometric fidelity of neural scene representations with the physical interpretability of modal analysis. Moreover, the proposed approach offers several technical advantages. The continuous representation inherent to NeRF enables high-resolution displacement tracking, critical for accurate modal parameter estimation. The combination of PCA and BSS provides robust performance in the presence of measurement noise and operational excitation. Furthermore, the framework naturally handles incomplete observations and occlusions through the learned scene priors embedded in the neural radiance field.

Acknowledgments This work was supported by CNPq (National Council for Scientific and Technological Development), Brazil (process no. 141761/2025-3).

References

1. Rainieri, C. "Perspectives of second-order blind identification for operational modal analysis of civil structures". *Shock Vib.*, 2014:1–9 (2014)ifnextchar.gobble.
2. Brincker, R. and Ventura, C. *Introduction to operational modal analysis*. John Wiley & Sons, Nashville, TN (2015)ifnextchar.gobble.
3. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., and Ng, R. "Nerf: representing scenes as neural radiance fields for view synthesis". volume 65, page 99–106, New York, NY, USA (2021)ifnextchar.gobble. Association for Computing Machinery.
4. Silva, M., Green, A., Morales, J., Meyerhofer, P., Yang, Y., Figueiredo, E., and Mascareñas, D. *Full-Field 3D Experimental Modal Analysis from Dynamic Point Clouds Measured Using a Time-of-Flight Imager*, pages 163–166 (2022)ifnextchar.gobble.
5. Poncelet, F., Kerschen, G., Golinval, J.C., and Marin, F. "System identification in structural dynamics using blind source separation techniques". In Davies, M.E., James, C.J., Abdallah, S.A., and Plumbley, M.D., editors, *Independent Component Analysis and Signal Separation*, pages 778–785, Berlin, Heidelberg (2007)ifnextchar.gobble. Springer Berlin Heidelberg.
6. Sadhu, A., Narasimhan, S., and Antoni, J. "A review of output-only structural mode identification literature employing blind source separation methods". *Mechanical Systems and Signal Processing*, 94:415–431 (2017)ifnextchar.gobble.
7. Zhang, T., Wang, C., and Zhang, Y. "Three-dimensional operational modal analysis based on self-iteration principal component extraction and direct matrix assembly". *Journal of Vibroengineering*, 19(8):6262–6276 (2017)ifnextchar.gobble.
8. Peeters, B. and Roeck, G.D. "Stochastic system identification for operational modal analysis: A review". *Journal of Dynamic Systems Measurement and Control-transactions of The Asme*, 123:659–667 (2001)ifnextchar.gobble.
9. Yang, Y. and Nagarajaiah, S. "Output-only modal identification by compressed sensing: Non-uniform low-rate random sampling". *Mechanical Systems and Signal Processing*, 56-57:15–34 (2015)ifnextchar.gobble.
10. Hu, K., Ying, W., Pan, Y., Kang, H., and Chen, C. "High-fidelity 3d reconstruction of plants using neural radiance fields". *Computers and Electronics in Agriculture*, 220:108848 (2024)ifnextchar.gobble. DOI:10.1016/j.compag.2024.108848.
11. Ribeiro, T., Brandão, F., Cardoso, V., Costa, J., Bittencourt, T., and Silva, M. "Neural radiance fields for low-cost, high-resolution 3d scanning of critical infrastructures" (2025)ifnextchar.gobble.
12. Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., Mcallister, D., Kerr, J., and Kanazawa, A. "Nerfstudio: A modular framework for neural radiance field development" (2023)ifnextchar.gobble.
13. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., and Hedman, P. " Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields ". In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469, Los Alamitos, CA, USA (2022)ifnextchar.gobble. IEEE Computer Society.
14. Zhou, L., Wu, G., Zuo, Y., Chen, X., and Hu, H. "A comprehensive review of vision-based 3d reconstruction methods". *Sensors*, 24(7) (2024)ifnextchar.gobble.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. "Attention is all you need" (2023)ifnextchar.gobble.
16. Kajiya, J. and von herzen, B. "Ray tracing volume densities". *ACM SIGGRAPH Computer Graphics*, 18:165–174 (1984)ifnextchar.gobble. DOI:10.1145/964965.808594.
17. Max, N. "Optical models for direct volume rendering". *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108 (1995)ifnextchar.gobble. DOI:10.1109/2945.468400.
18. Kazemnejad, A., Padhi, I., Ramamurthy, K.N., Das, P., and Reddy, S. "The impact of positional encoding on length generalization in transformers" (2023)ifnextchar.gobble.
19. Pepino, L., Riera, P., and Ferrer, L. "Study of positional encoding approaches for audio spectrogram transformers". In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 3713–3717. IEEE (2022)ifnextchar.gobble. DOI:10.1109/icassp43922.2022.9747742.
20. Cantürk, S., Liu, R., Lapointe-Gagné, O., Létourneau, V., Wolf, G., Beaini, D., and Rampášek, L. "Graph positional and structural encoder" (2024)ifnextchar.gobble.
21. Terven, J., Cordova-Esparza, D.M., Ramirez-Pedraza, A., Chavez-Urbiola, E.A., and Romero-Gonzalez, J.A. "Loss functions and metrics in deep learning" (2024)ifnextchar.gobble.

22. Wang, Q., Ma, Y., Zhao, K., and Tian, Y. "A comprehensive survey of loss functions in machine learning". *Annals of Data Science*, 9(2):187–212 (2020)ifnextchar.gobble. DOI:10.1007/s40745-020-00253-5.
23. Zhang, T., Zhu, D., Zhang, G., Shi, W., Liu, Y., Zhang, X., and Li, J. "Spatiotemporally enhanced photometric loss for self-supervised monocular depth estimation". In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE (2022)ifnextchar.gobble.
24. Zhao, H., Gallo, O., Frosio, I., and Kautz, J. "Loss functions for image restoration with neural networks". *IEEE Transactions on Computational Imaging*, 3(1):47–57 (2017)ifnextchar.gobble. DOI:10.1109/TCI.2016.2644865.
25. Yang, Y. and Nagarajaiah, S. "Blind modal identification of output-only structures in time-domain based on complexity pursuit". *Earthquake Engineering & Structural Dynamics*, 42(13):1885–1905 (2013)ifnextchar.gobble.
26. Yang, Y. and Nagarajaiah, S. "Structural damage identification via a combination of blind feature extraction and sparse representation classification". *Mechanical Systems and Signal Processing*, 45(1):1–23 (2014)ifnextchar.gobble.
27. Yang, Y., Dorn, C., Mancini, T., Talken, Z., Kenyon, G., Farrar, C., and Mascareñas, D. "Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification". *Mechanical Systems and Signal Processing*, 85:567–590 (2017)ifnextchar.gobble.
28. Golub, G.H. and Van Loan, C.F. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, USA (1996)ifnextchar.gobble.
29. Feeny, B. and Kappagantu, R. "On the physical interpretation of proper orthogonal modes in vibrations". *Journal of Sound and Vibration*, 211(4):607–616 (1998)ifnextchar.gobble.
30. Stone, J.V. "Blind source separation using temporal predictability". *Neural Computation*, 13(7):1559–1574 (2001)ifnextchar.gobble.