2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Metric, inertially aligned monocular state estimation via kinetodynamic priors

Jiaxin Liu[1*], Min Li[1*], Wanting Xu[1], Liang Li[2], Jiaqi Yang[1] and Laurent Kneip[1]

***Abstract*— Accurate state estimation for flexible robotic systems poses significant challenges, particularly for platforms with dynamically deforming structures that invalidate rigid-body assumptions. This paper addresses this problem and enables the extension of existing rigid-body pose estimation methods to non-rigid systems. Our approach integrates two core components: first, we capture elastic properties using a deformation-force model, efficiently learned via a Multi-Layer Perceptron; second, we resolve the platform's inherently smooth motion using continuous-time B-spline kinematic models. By continuously applying Newton's Second Law, our method formulates the relationship between visually-derived trajectory acceleration and predicted deformation-induced acceleration. We demonstrate that our approach not only enables robust and accurate pose estimation on non-rigid platforms, but also shows that the properly modeled platform physics allow for the recovery of inertial sensing properties. We validate this feasibility on a simple spring-camera system, showing how it robustly resolves the typically ill-posed problem of metric scale and gravity recovery in monocular visual odometry.**

I. Introduction

Accurate environmental perception and precise self-pose estimation are paramount for autonomous navigation, human-robot collaboration, and complex robotic tasks. Traditionally, these capabilities have relied on rigid-body assumptions, simplifying multi-sensor fusion and motion estimation. However, the burgeoning field of soft robotics and flexible systems—such as compliant manipulators and UAVs—challenges this paradigm [1]. While valued for their inherent safety, enhanced dexterity, and potential for lower manufacturing costs, these non-rigid systems leverage structural deformation for environmental adaptation, which critically introduces dynamic, time-variant relative sensor poses (Fig. 1). This invalidates classical rigid-body algorithms, presenting significant perception and state estimation challenges [2]. In this work, we demonstrate that non-rigid platforms may not necessarily complicate state estimation and sensor fusion. Rather, we show how non-rigid elements and kinetodynamic priors [3] can even add additional constraints to the system. The leading example we use is monocular motion estimation. While the recovery of scale and inertial alignment are severely ill-posed problems that would normally require fusion with additional sensors [4], [5], the exploitation of motion and deformation priors perhaps surprisingly renders these dimensions observable.

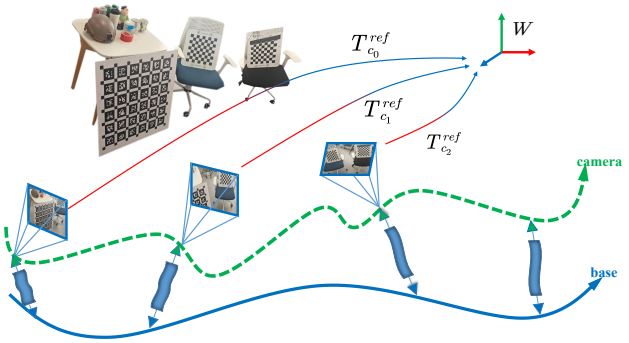

Fig. 1: Non-rigid monocular camera system demonstrating trajectory divergence. Unlike rigid systems, the deformable connection induces camera oscillations (green) distinct from the platform's motion (blue).

This work draws inspiration from the work of Li et al. [6], which demonstrated the successful embedding of a physical deformation model into a non-rigid multi-perspective camera system. Intriguingly, the deformable connections between sensors in that work acted as a "passive IMU". Building on this insight, but targeting robust state estimation using a single exteroceptive sensor, this paper addresses a key limitation of prior models. We adopt a mechanical design similar to the Zebedee system [7], where a monocular camera is connected to a mobile platform via a spring mechanism. While this setup serves as a scientific example to explore the potential of priors in non-rigid platforms, we note that there are practically used multi-sensor heads in which non-rigidity was actively added to, for example, enlarge the perceptive field of view [8], [9].

To robustify state estimation under such dynamic non-rigid motion, we explicitly incorporate kinetodynamic priors. This work introduces a novel framework that unifies kinematic and dynamic constraints for non-rigid systems, built upon two core assumptions: 1) Continuous-time Kinematic Models: We use B-Splines to model the smooth motion of at least one platform point, enabling the derivation of high-order derivatives crucial for dynamic analysis. 2) Learned Deformation-force Model: The platform's elastic properties are captured through an injective deformation-force model, efficiently learned via a Multi-Layer Perceptron [10], thus bypassing computationally expensive Finite Element Analysis [11]. This approach resolves critical challenges like scale and gravity recovery in monocular visual odometry. It continuously applies Newton's Second Law, establishing a

*Two authors contribute equally in this work

[1]Jiaxin Liu, Min Li, Wanting Xu, Jiaqi Yang and Laurent Kneip are with School of Information Science and Technology, ShanghaiTech University, Shanghai, China. {liujx2024, limin1, xuwt, yangjq, lkneip}@shanghaitech.edu.cn

[2]Liang Li is with the Department of Mechanical Engineering, The University of Hongkong, Hongkong S.A.R., China. llihku@connect.hku.hk

Corresponding author: Laurent Kneip, lkneip@shanghaitech.edu.cn

physical relationship between the camera's visually derived trajectory acceleration and the acceleration predicted by our learned deformation-force model. Minimizing the discrepancy between these accelerations allows us to determine the unknown scale factor, enabling metric pose and scale estimation from monocular vision [12]. Effectively, any camera motion not explained by the smooth body trajectory is attributed to spring deformations, thus aligning with the support structure's physical properties through an inertial aligning transformation to achieve metric and inertially aligned monocular ego-motion estimation.

Our contributions are summarized as follows:

- We introduce compact neural representations for modeling elastic deformation properties of sensor support platforms, coupled with a calibration method using a motion capture device.
- We demonstrate how the combination of a suitable body motion model and an elastic deformation model can be leveraged for passive inertial sensing and accurate monocular exteroceptive motion estimation in non-rigid scenarios.
- We present the complete computational paradigm, encompassing numerical differentiation of the camera trajectory, variable initialization, and an optimization framework with an embedded, differentiable neural body deformation model.

As demonstrated by our results, this outlined setup develops inherent passive inertial sensing capabilities, demonstrating the feasibility of accurate real-world motion estimation from ego-centric video through a mostly mathematical model. Although tested on an exemplary setup, we believe this approach holds significant promise and is applicable to a wide range of future robotic platforms possessing specific motion models and potentially elastic actuation chains.

II. Related Work

A. State Estimation on Non-Rigid Platforms

The integration of flexible elements into robotic systems, such as deformable UAV wings [13] and soft robots [14], [15], [16], [17], [18], has gained significant traction. This paradigm contrasts sharply with traditional robotics, which assumes sensors follow rigid trajectories. The dynamic changes in sensor-to-body transformations pose a significant challenge to conventional state estimation algorithms that rely on rigid-body assumptions. A few exceptions exist that explicitly model non-rigid systems, such as the work by Peng et al. [19] and Hinzmann et al. [14], [15]. However, these methods are either limited to non-elastic [19] or static scenarios [6] or depend on overlapping sensor fields of view [15]. Our work, in contrast, tackles the more challenging problem of pure monocular visual state estimation on a non-rigid platform in dynamic environments. The Zebedee system [7] is a seminal example that leverages a platform's elastic motion to extend sensor viewpoints, thereby achieving large-scale 3D coverage, but our approach focuses on using the non-rigidity itself as a source of information.

B. Monocular Visual and Multi-Sensor State Estimation

Monocular cameras are simple and cheap, but they suffer from a fundamental limitation: scale ambiguity. Classic methods like ORB-SLAM [20], [21], [22] and COLMAP/GLOMAP [23], [24], [25], [26] can reconstruct 3D maps and trajectories, but their scale is always relative. This means the reconstructed environment can be resized by an unknown factor, making it unsuitable for tasks requiring absolute measurements. To overcome this, many systems fuse a camera with other sensors. For example, VINS-Mono [27] and maplab [28] combine a camera with an IMU to resolve scale using inertial data. Similarly, other methods use LiDAR [29], [30], [31] or GPS [32]. While effective, these solutions increase hardware cost and complexity. Recently, learning-based methods like Mast3r-SLAM [33] and VGGt-SLAM [34] have tried to predict scale directly from images. However, these often require significant computation and cannot be used in a real-time system.

Instead of adding new sensors, we use the non-rigid platform's own elastic properties to resolve scale ambiguity. By modeling the deformation, we achieve metrically accurate state estimation with just a single camera, providing a novel solution for flexible robotics without additional hardware.

C. Passive Sensing and Kinetodynamic Modeling

Our work is inspired by the concept of "passive inertial sensing" from Li et al. [6], who demonstrated that a sensor's support deformation under gravity can be used to infer inertial information. However, their method was limited to a specific, static physical model. To overcome the limitations of traditional physical models, we employ a more general, data-driven neural model to characterize complex, continuous deformations. While neural networks have been applied to model soft robot control [35] and simulate soft body deformations [36], these works primarily focus on feed-forward or control problems. In contrast, our approach innovatively integrates a differentiable neural representation of the non-rigid connection's physical properties directly into a monocular visual state estimation framework. This allows us to perform passive perception of continuous deformation states, offering a new perspective for flexible robotics and embodied intelligence research.

III. Methodology

Fig. 2 illustrates our pipeline for motion recovery in a system comprising a camera c and a moving platform b joined by a non-rigid connection. Our approach addresses the coupling of these components in two stages: first, we characterize the elastic link using a Deformation-force Network (DFN); second, we leverage the learned physics prior to recover the metric trajectory and scale of the platform via a B-Spline-based joint optimization.

Throughout this paper, subscripts are used to denote the specific entity (e.g., c for the camera, b for the base), while superscripts indicate the reference coordinate frame in which a quantity is expressed.

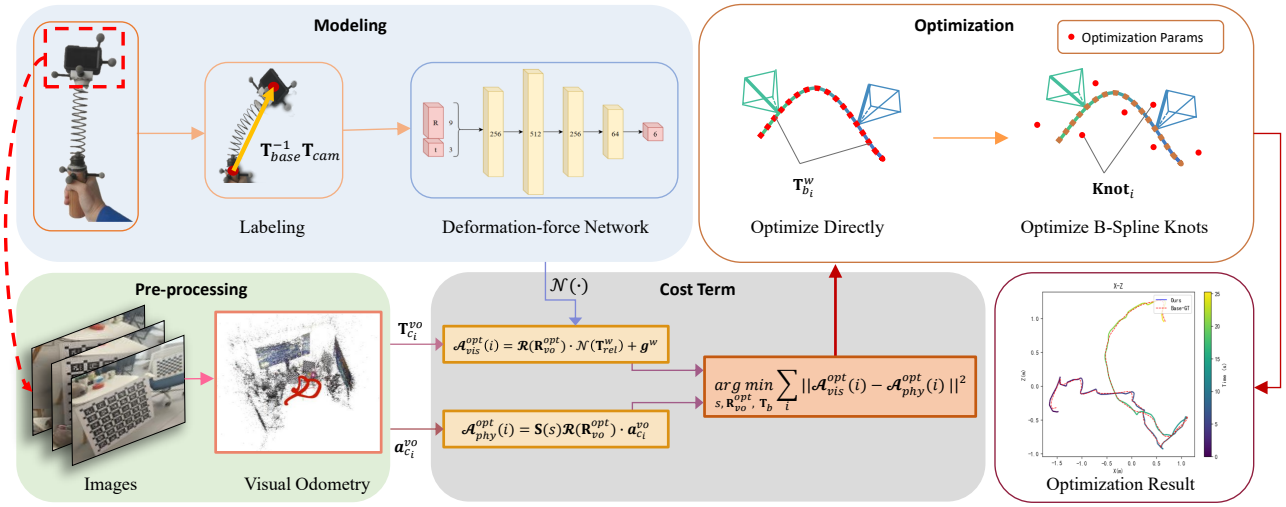

Fig. 2: Overview of the proposed pipeline. **a) Modeling:** A DFN is trained offline to implicitly characterize the non-rigid connection using linear and angular accelerations. **b) Solving:** We first recover the camera trajectory via visual odometry, followed by a B-Spline-based optimization that minimizes the discrepancy between visual motion and learned physics to estimate the platform's metric state.

A. Preliminaries: B-Spline for $\mathbb{SE}(3)$

For a k-th order B-Spline with control points $\mathbf{T}_i, \mathbf{T}_{i+1}, \ldots, \mathbf{T}_N$, the spline is defined as:

$$\mathbf{T}(t) = \mathbf{T}_i \prod_{j=1}^{k-1} \mathrm{Exp}(\widetilde{\mathbf{B}}_j(t) \cdot \mathrm{Log}(\mathbf{T}_{i+j-1}^{-1} \cdot \mathbf{T}_{i+j})), \tag{1}$$

where $\mathrm{Exp}(\cdot)$ and $\mathrm{Log}(\cdot)$ represent the exponential and logarithmic mappings on $\mathbb{SE}(3)$, respectively. The basis functions $\widetilde{\mathbf{B}}_j(t)$ are determined by the blending matrix $\widetilde{\mathbf{M}}$ with entries:

$$m_{s,n}^{(k)} = \frac{C_{k-1}^{n}}{(k-1)!} \sum_{l=s}^{k-1} (-1)^{l-s} C_k^{l-s} (k-1-l)^{k-1-n} \qquad s,n \in 0,..,k-1. \tag{2}$$

The velocity can be derived by defining $\mathbf{A}_j(t) = \mathrm{Exp}(\widetilde{\mathbf{B}}_j(t))\mathrm{Log}(\mathbf{T}_{i+j-1}^{-1} \cdot \mathbf{T}_{i+j})$, leading to:

$$\dot{\mathbf{T}}(t) = \mathbf{T}_i \cdot \sum_{j=1}^{k-1} \left(\prod_{l=1}^{j-1} \mathbf{A}_l(t) \right) \dot{\mathbf{A}}_j(t) \left(\prod_{l=j+1}^{k-1} \mathbf{A}_l(t) \right). \tag{3}$$

Eq. 3 yields the linear and angular velocities from the visual odometry trajectory. The corresponding accelerations can be derived in a similar manner by taking the second time derivative, both of which are subsequently used to formulate the optimization objectives.

B. Modeling: Deformation-force Network

The mechanical properties of a non-rigid connection can be modeled using a nonlinear differential equation with damping characteristics [37]:

$$\boldsymbol{f}_s = \kappa_1 \Delta \boldsymbol{x} + \kappa_3 \Delta \boldsymbol{x}^3 + c \frac{d\Delta \boldsymbol{x}}{dt}, \tag{4}$$

where $\boldsymbol{f}_s$ is the elastic force exerted by the non-rigid connection on the camera. Neglecting air resistance and potential collisions under low-speed conditions, the camera's dynamics are governed by Newton's second law. The relationship between the camera's linear and angular accelerations and its relative displacement is implicitly modeled as:

$$\boldsymbol{a}_c - \boldsymbol{g} = H(\Delta \boldsymbol{x}, \Delta \boldsymbol{\theta}), \tag{5}$$

where $\boldsymbol{a}_c = \begin{bmatrix} \boldsymbol{a}_c^{linear^T} & \boldsymbol{a}_c^{angular^T} \end{bmatrix}^T$ is the 6-DoF acceleration and $\boldsymbol{g} = [0,0,-g,0,0,0]^T$ represents gravity. The relative pose $\mathbf{T}_b^{-1}\mathbf{T}_c$ captures the structural deformation. Thus, we approximate this mapping using a neural network $\mathcal{N}$:

$$\boldsymbol{a}_c - \boldsymbol{g} = \mathcal{N}(\mathbf{T}_b^{-1}\mathbf{T}_c). \tag{6}$$

To ensure the network learns physics in a consistent sensor-centric frame, we supervise it using ground-truth motion projected into the camera frame c:

$$\boldsymbol{a}_c^c - \boldsymbol{g}^c = \mathcal{R}(\mathbf{R}_c^{gt_i})^{-1}(\boldsymbol{a}_c^{gt_i} - \boldsymbol{g}_c^{gt_i}), \tag{7}$$

where $\mathcal{R}(\mathbf{R}) = \mathrm{diag}(\mathbf{R}, \mathbf{R})$ is a 6×6 block-diagonal rotation matrix, and gt_i denotes the ground-truth inertial frame. The ground-truth kinematic data is captured via an external motion capture system and is exclusively utilized during the offline training phase.

C. Optimization: State Estimation and Scale Recovery

Visual odometry (VO) provides unscaled camera trajectories. We relate these to the optimized metric trajectories via a rotation $\mathbf{R}_{vo}^{opt}$ and a scaling factor s. Since angular motion is scale-invariant, we define the scaling matrix $\mathbf{S}(s) = \mathrm{diag}(s,s,s,1,1,1)$. The visual acceleration constraint is thus:

$$\boldsymbol{a}_{c_i}^{opt} = \mathbf{S}(s)\mathcal{R}(\mathbf{R}_{vo}^{opt})\boldsymbol{a}_{c_i}^{vo}. \tag{8}$$

Physically, the optimized acceleration $\boldsymbol{a}_{c_i}^{opt}$ must also satisfy the dynamics predicted by the network. By transforming the network output to the *opt* frame, we obtain:

$$\boldsymbol{a}_{c_i}^{opt} = \mathcal{R}(\mathbf{R}_{c_i}^{opt})\mathcal{N}((\mathbf{T}_{b_i}^{opt})^{-1}\mathbf{T}_{c_i}^{opt}) + \boldsymbol{g}^{opt}. \tag{9}$$

Defining the residual $\boldsymbol{r}_i$ as the discrepancy between physical prediction and visual observation:

$$\boldsymbol{r}_i = \mathcal{R}(\mathbf{R}_{c_i}^{opt})\mathcal{N}((\mathbf{T}_{b_i}^{opt})^{-1}\mathbf{T}_{c_i}^{opt}) + \boldsymbol{g}^{opt} - \mathbf{S}(s)\mathcal{R}(\mathbf{R}_{vo}^{opt})\boldsymbol{a}_{c_i}^{vo}. \tag{10}$$

Here, we directly set $\boldsymbol{g}^{opt} = \boldsymbol{g}^{w}$ and then verify $\mathbf{R}_{vo}^{opt}$ to ensure the angle between the optimized coordinate system and the ground truth gravity direction remains small.

To ensure efficiency and smoothness, we parameterize the platform trajectory using B-Spline control points $\mathbf{T}_i^{opt}$. The joint optimization objective is:

$$\underset{s,\mathbf{T}_{0:n}^{opt},\mathbf{R}_{vo}^{opt}}{\arg\min} \sum_i \|\boldsymbol{r}_i\|^2. \tag{11}$$

This formulation enables the recovery of the metric scale s by aligning the dimensionless visual motion with the metric forces learned from structural deformations.

IV. Experiments

A. Implementation Details

(a) (b) (c)

Fig. 3: Hardware setup. The system comprises a base, a camera, and a non-rigid spring connection. The optimization targets are the pose and scale of the base.

1) Hardware and Data Acquisition: As illustrated in Fig. 3, the experimental non-rigid system consists of a monocular camera attached to a moving base via a passive elastic spring. To capture irregular motion induced by body-induced accelerations and vibrations, a handheld holder is utilized. The camera's 6-DoF motion is governed by the interaction between the spring's restoration force and gravity.

Ground truth trajectories are acquired using an optical motion capture system. Markers are attached to both the camera and the holder (base). The coordinate systems are gravity-aligned to ensure consistency with the gravity-aware network training and optimization. Trajectory sequences, ranging from 25 to 45 seconds, were recorded for evaluation.

2) Optimization Setup: We employ COLMAP [23] for robust visual odometry. As described in Eq. 11, the optimization terms for the base trajectory and scale are constructed using the camera's trajectory and acceleration derived from VO. The problem is solved using the Ceres Solver.

3) Evaluation Metrics: Performance is evaluated using three key metrics:

- **Trajectory Accuracy:** We compute the Absolute Pose Error (APE) for both the VO output ($\mathbf{T}_{c_i}^{vo}$ vs. $\mathbf{T}_{c_i}^{gt}$) and the optimized base trajectory ($\mathbf{T}_{b_i}^{w}$ vs. $\mathbf{T}_{b_i}^{gt}$).
- **Scale Accuracy:** The relative scale error is defined as $err_s = \frac{\|s_{gt}-s\|}{s_{gt}}$, where s_{gt} is the ground truth scale aligned using EVO [38].
- **Gravity Alignment:** Since physical gravity is integrated into our pipeline, we evaluate the angular error between the optimized gravity direction and the physical gravity vector: $err_G = \arccos(\frac{\boldsymbol{g}^{opt}\cdot\boldsymbol{g}^{gt}}{\|\boldsymbol{g}^{opt}\|\cdot\|\boldsymbol{g}^{gt}\|})$.

B. Experimental Results

1) Simulation Experiments: To assess robustness against noise and outliers, we generated synthetic camera trajectories by applying Gaussian noise and outliers to real-world ground truth data. The simulated pose $\mathbf{T}_{c_i}^{sim}$ is defined as:

$$\mathbf{T}_{c_i}^{sim} = \mathbf{T}_{c_i}^{gt} \oplus (\sigma \cdot \boldsymbol{\xi}), \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{12}$$

where σ represents the noise amplitude. Additionally, to simulate gross errors, a proportion of the trajectory corresponding to the outlier ratio $\mathbf{O}_r$ is replaced with random transformations. We conducted experiments across varying noise levels (0% to 10%) and outlier ratios (0% to 5%). For each setting, results were averaged over six independent runs.

TABLE I: Average performance metrics under varying noise magnitudes.

Noise	APE			err_s	err_G (°)
	Mean (m)	Median (m)	STD		
0%	0.036	0.035	0.014	0.006	0.447
3%	0.040	0.038	0.043	0.003	0.638
5%	0.067	0.062	0.034	0.058	1.328
10%	0.096	0.092	0.041	0.042	1.841

TABLE II: Average performance metrics under varying outlier proportions (fixed 3% noise).

Outlier	APE			err_s	err_G (°)
	Mean (m)	Median (m)	STD		
0%	0.040	0.038	0.043	0.003	0.638
1%	0.083	0.080	0.039	0.062	1.579
3%	0.141	0.138	0.042	0.065	1.943
5%	0.164	0.158	0.061	0.108	2.657

Table I demonstrates the method's robustness to amplitude noise. Even at a 10% noise level, scale and gravity errors remain low. Furthermore, as shown in Table II, the algorithm maintains acceptable accuracy even when challenged with up to 5% outliers, validating its stability in non-ideal conditions.

2) Real-World Experiments: In real-world scenarios, we utilize the estimated $\mathbf{T}_{c_i}^{vo}$ and acceleration $\mathbf{a}_{c_i}^{vo}$ from COLMAP to optimize the base trajectory. In our framework, Visual Odometry is treated as an input component, providing initial

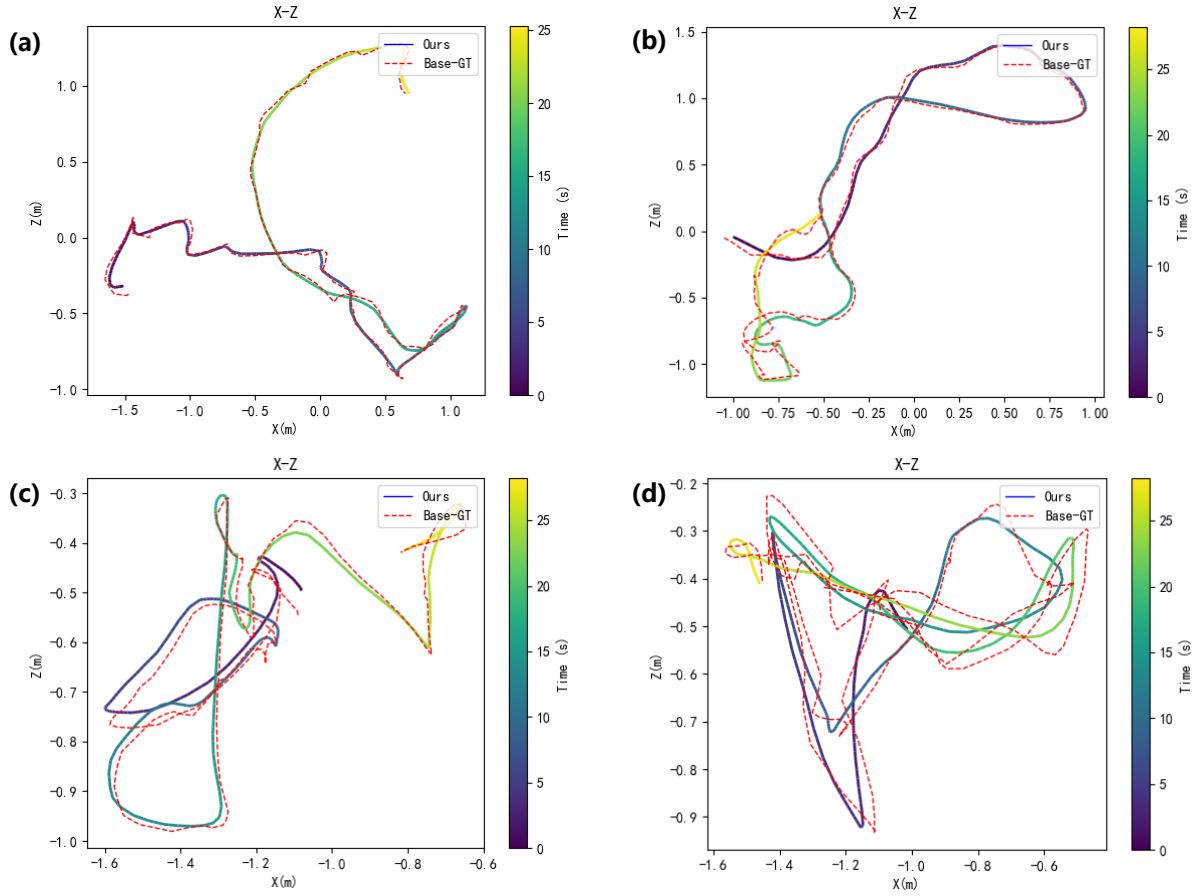

Fig. 4: Qualitative results. 2D X-Z projections of ground truth (solid) and optimized (dashed) trajectories are shown. The alignment demonstrates the effectiveness of the proposed method across various motion patterns. (d) illustrates a challenging case with motion blur.

kinematic constraints for the subsequent optimization. To the best of our knowledge, no other established methods effectively address trajectory estimation specifically for such non-rigid systems.

Fig. 4 visualizes the optimization results. The strong overlap between estimated and ground truth trajectories confirms the accuracy of our pipeline. However, large system deformations can induce rapid camera motion, causing motion blur that degrades VO performance. This is observed in Fig. 4(d), where accuracy is slightly reduced compared to other sequences. This limitation could be mitigated by using high-frame-rate cameras or ensuring a feature-rich environment.

Quantitative results are detailed in Table III. The analysis confirms that our method successfully recovers the metric scale and estimates the base trajectory of a non-rigid system using only a monocular camera, without requiring additional sensors.

C. Discussion

The performance discrepancy between simulation and real-world experiments primarily stems from the quality of the input camera trajectory. While simulations utilize ground truth with additive Gaussian noise to isolate algorithmic efficacy, real-world experiments rely on visual odometry for state estimation. Crucially, the non-rigid spring coupling induces high-frequency vibrations and rapid camera motion, resulting in significant motion blur. This degradation in image quality inevitably impairs the VO tracking accuracy. Since our proposed pipeline depends on the estimated camera kinematics, errors from the VO stage propagate to the final base trajectory optimization, thereby limiting the overall system precision in dynamic real-world scenarios compared to the idealized simulation environment.

V. Ablation Study

To validate the importance of our data collection and processing pipeline for accurately modeling a non-rigid system, we design two ablation studies. In the first study, we eliminate the step of normalizing data to the camera coordinate system. The second study demonstrates the advantage of using multi-dimensional motion patterns for robustly modeling gravity.

1) Relative Pose Normalization (Eq. 7): To demonstrate the importance of normalizing data to the camera coordinate system, we reconstructed the dataset and performed system identification again, this time removing the step represented

TABLE III: Quantitative error analysis on real-world experimental data. APE (Optimization) measures the error between optimized base pose $\mathbf{T}_{b_i}^{w}$ and ground truth $\mathbf{T}_{b_i}^{gt}$; APE (VO Component) measures the error between visual odometry $\mathbf{T}_{c_i}^{vo}$ and ground truth $\mathbf{T}_{c_i}^{gt}$. Note that the VO trajectory is Sim(3)-aligned to the ground truth. Utilizing this oracle scale may occasionally make VO translation errors lower than our optimized results.

Seq.	APE (Optimization)			APE (VO Component)			err_s	s	s_{gt}	Gravity Err. (°)
	Mean(m)	Median(m)	STD	Mean (m)	Median (m)	STD				
1	0.120	0.081	0.088	0.196	0.201	0.086	0.271	0.183	0.251	7.34
2	0.149	0.138	0.069	0.158	0.147	0.086	0.246	0.308	0.247	6.67
3	0.170	0.121	0.039	0.159	0.154	0.086	0.201	0.194	0.243	4.83
4	0.136	0.101	0.068	0.274	0.273	0.156	0.147	0.132	0.115	3.28
5	0.181	0.188	0.093	0.127	0.117	0.077	0.276	0.143	0.112	6.28
6	0.165	0.172	0.064	0.161	0.132	0.117	0.707	0.031	0.106	13.4
7	0.158	0.152	0.081	0.234	0.212	0.112	0.669	0.202	0.121	6.88
8	0.150	0.156	0.081	0.306	0.251	0.211	0.170	0.112	0.135	7.10
9	0.283	0.257	0.148	0.289	0.255	0.165	0.622	0.318	0.196	6.82
10	0.213	0.154	0.143	0.268	0.238	0.141	0.711	0.344	0.201	8.22
11	0.218	0.211	0.094	0.292	0.293	0.155	0.372	0.177	0.129	7.12
12	0.293	0.254	0.142	0.248	0.240	0.128	0.476	0.056	0.107	2.12
13	0.715	0.598	0.406	0.563	0.584	0.217	1.509	0.527	0.210	5.98
14	0.120	0.081	0.088	0.121	0.110	0.072	0.648	0.183	0.521	7.34
15	0.461	0.399	0.307	0.285	0.294	0.171	0.027	0.258	0.251	0.92
16	0.094	0.084	0.042	0.096	0.095	0.045	0.671	0.028	0.085	7.27
Median	0.167	0.155	-	0.241	0.225	-	0.483	-	-	6.85
Mean	0.226	0.196	-	0.236	0.224	-	0.424	-	-	6.36

TABLE IV: Comparison of errors: Normalized (with Eq. 7) vs. Directly (without Eq. 7). We specifically report metrics such as the mean and median errors for linear and angular accelerations across each axis.

		Mean	Median	STD
$E_{normalized}$	${}^{(x)}a_c^{linear}$ (m/s^2)	0.323	0.327	0.151
	${}^{(y)}a_c^{linear}$ (m/s^2)	0.799	0.653	0.636
	${}^{(z)}a_c^{linear}$ (m/s^2)	0.167	0.147	0.109
	${}^{(x)}a_c^{angular}$ (rad/s^2)	0.148	0.119	0.122
	${}^{(y)}a_c^{angular}$ (rad/s^2)	0.121	0.092	0.109
	${}^{(z)}a_c^{angular}$ (rad/s^2)	0.203	0.168	0.171
$E_{directly}$	${}^{(x)}a_c^{linear}$ (m/s^2)	1.128	0.91	0.898
	${}^{(y)}a_c^{linear}$ (m/s^2)	1.578	1.197	1.213
	${}^{(z)}a_c^{linear}$ (m/s^2)	1.251	1.028	1.089
	${}^{(x)}a_c^{angular}$ (rad/s^2)	0.456	0.401	0.338
	${}^{(y)}a_c^{angular}$ (rad/s^2)	0.353	0.296	0.283
	${}^{(z)}a_c^{angular}$ (rad/s^2)	0.405	0.377	0.261

by Eq. 7. Table IV shows the resulting modeling errors, listing the acceleration and angular acceleration errors in each axis. These results highlight the significant advantage of our normalization approach.

2) Movement Patterns on Modeling: As previously mentioned, we deliberately curated a diverse set of motion patterns for our training data. This strategy was employed to ensure the final model could accurately handle the constant-direction gravitational vector. Table V presents the results of training with these distinct patterns. The results indicate that while individual patterns enable the network to learn dynamics specific to certain axes, a combination of all patterns is crucial for the network to generalize and robustly handle all motion scenarios.

TABLE V: Modeling performance with different motion patterns and the units are m/s^2 and rad/s^2, respectively. The best and second-best performing metrics are bolded for clarity. (Pattern A: Pure translation, B: Pure rotation, C: Translation + vertical, D: Rotation + vertical, detailed in Appendix A.)

	Pattern A	Pattern B	Pattern C	Pattern D	Total
${}^{(x)}a_c^{linear}$	1.071	**0.864**	1.198	1.041	**0.939**
${}^{(y)}a_c^{linear}$	5.244	4.816	4.808	**4.186**	**2.325**
${}^{(z)}a_c^{linear}$	**0.534**	1.781	0.665	0.834	**0.485**
${}^{(x)}a_c^{angular}$	**0.368**	0.375	**0.348**	0.400	0.402
${}^{(y)}a_c^{angular}$	0.376	**0.355**	0.455	0.436	**0.329**
${}^{(z)}a_c^{angular}$	**0.346**	**0.353**	0.449	0.447	0.552

VI. Conclusion and Discussions

This paper introduced a novel perception system leveraging a monocular camera passively coupled to a moving platform via a non-rigid connection. By integrating a learned deformation-force model with our state estimation framework, we demonstrated that non-rigid coupling provides crucial constraints for passive inertial sensing, effectively resolving inherent scale and inertial alignment ambiguities. Our experiments validated that correctly modeling kinetodynamics enables accurate motion and metric scale estimation.

A fundamental limitation of our system is its reliance on the coupling's elasticity; as the connection approaches ideal rigidity, the informative deformation cues diminish, and the system's advantage degenerates to standard monocular setups. Additionally, the current batch optimization framework entails increasing computational overhead over long

trajectories. We also observed that rotational precision, while competitive, remains more sensitive to optimization on the $\mathbb{SO}(3)$ manifold compared to translation. Future work will focus on implementing a sliding-window optimization for real-time efficiency and investigating manifold-aware loss functions to further refine rotational accuracy. This approach marks a significant step toward robust state estimation for future robotic platforms with elastic actuation chains.

VII. Acknowledgement

The authors would like to thank the fund support from Natural Science Foundation of China(W2531052) and the 3D printing support from ShanghaiTech SIST Machine Shop on real experiment setup.

Appendix

A. Data Acquisition

We use a tracking system to acquire ground truth trajectories of both the camera and the base at 360 Hz. We collected trajectories lasting 25-45 seconds across various patterns, encompassing both smooth and vigorous motions with diverse poses. As illustrated in Fig. A1, for each pose, we collected data for:

- **Pure translational motion**: The spring primarily undergoes tangential deformation along the direction of motion.
- **Pure rotational motion**: The camera experiences radial stretching of the spring due to centrifugal force.
- **Translational motion with vertical movement (gravity aligned)**: This setup amplifies the effect of gravity, enabling accurate modeling of its influence.
- **Rotational motion with vertical movement (gravity aligned)**: This case induces the most complex spring deformation, resulting from the combined effects of centrifugal force and gravity.

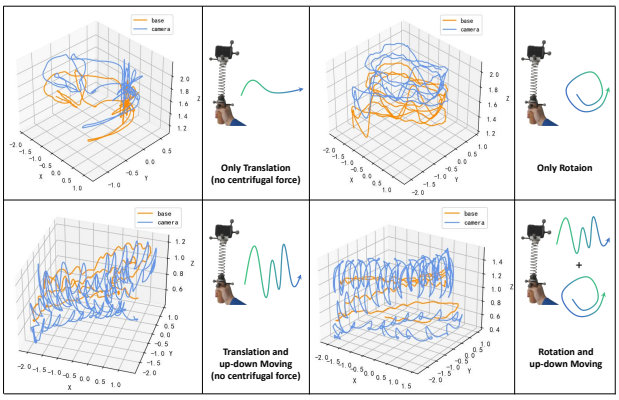

Fig. A1: Trajectories for different states of motion.

To capture the spring's behavior under diverse gravitational influences, we collected data with the system inverted and on its side, addressing the compressive effect of gravity in a vertical downward orientation. Fig. A2 illustrates these varied configurations, alongside trajectories demonstrating different intensities of complex motion states.

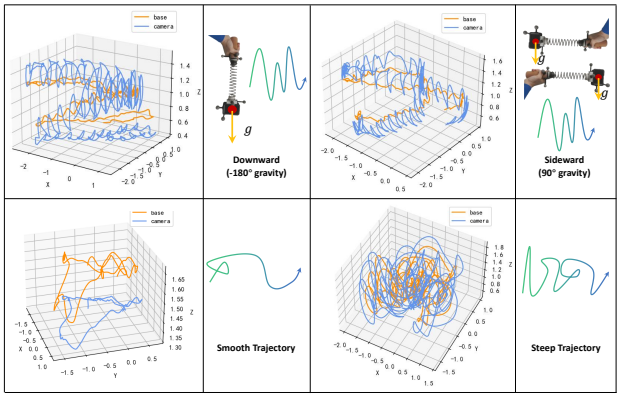

Fig. A2: Trajectories with varying gravitational orientations and smoothness levels.

B. Training DFN

We employ an MLP to model the non-rigid connection. The network's computational complexity and parameter count were quantified using the `thop` library, revealing 375.392 KFLOPs and 377.222 K parameters, respectively.

We use approximately 259 K samples derived from 16 collected trajectories for training, and another 16 sequences reserved for real experiments. These samples are partitioned into training, testing, and validation sets in a 7:2:1 ratio. We use Adam as an optimizer, and the learning rate is 1e-4. We train the network for 100 epochs with the batch size 1024.

C. Observability Analysis of the Metric Scale

The metric scale s is inherently unobservable in pure monocular visual odometry due to the scale-invariant nature of perspective projection. In our framework, this scale ambiguity is resolved through the macroscopic coupling of visual kinematics and learned physical dynamics.

Let $\mathbf{a}_{ci}^{vo,linear}$ be the dimensionless linear acceleration derived from visual odometry. The physical consistency residual for the linear acceleration component can be expressed as a function of the scale s:

$$\mathbf{r}_i^{linear}(s) = \mathbf{R}_{ci}^{opt}\,\mathscr{N}_{linear}\left(\mathbf{T}_{c_i}^{b_i}(s)\right) + \mathbf{g}^w - s\mathbf{R}_{vo}^{opt}\,\mathbf{a}_{ci}^{vo,linear}, \quad \text{(C1)}$$

where $\mathscr{N}_{linear}(\cdot)$ extracts the linear acceleration prediction from the Deformation-force Network, and $\mathbf{g}^w$ represents the physical gravity vector.

The network $\mathscr{N}$ is trained offline using ground-truth kinematics captured by a motion capture system. This methodology ensures that $\mathscr{N}$ intrinsically operates in a true metric space. It acts as a learned kinetodynamic prior, mapping relative structural deformations directly to absolute metric forces and accelerations, thereby embedding the physical elasticity scale into the model's weights.

And the formulation explicitly incorporates the physical gravity vector $\mathbf{g}^w$ as an absolute metric reference, which inherently prevents any potential linear scaling equivalence between the kinematics and the dynamics. Specifically, if an erroneous, arbitrary scaling factor $\alpha \neq 1$ were applied to the spatial dimensions of the system, the purely kinematic visual acceleration would scale proportionally. However, the

physical acceleration prediction would not scale equivalently due to the constant additive gravity term:

$$\alpha s \mathbf{R}_{vo}^{opt} \mathbf{a}_{ci}^{vo,linear} \neq \mathbf{R}_{ci}^{opt} \mathscr{N}_{linear} \left(\mathbf{T}_{c_i}^{b_i}(\alpha s) \right) + \mathbf{g}^{w}. \quad \text{(C2)}$$

This inequality demonstrates that an arbitrary scale cannot arbitrarily cancel out across the dynamic constraints. As a result, the optimal scale s is uniquely determined and aligned to the absolute metric scale defined by the gravitational constant.

REFERENCES

[1] Z. Chen, D. Wu, Q. Guan, D. Hardman, F. Renda, J. Hughes, T. G. Thuruthel, C. Della Santina, B. Mazzolai, H. Zhao, *et al.*, "A Survey on Soft Robot Adaptability: Implementations, Applications, and Prospects [Survey]," *IEEE Robotics & Automation Magazine*, 2025.

[2] H. P. Thanabalan, "Learning Soft Robot and Soft Actuator Dynamics using Deep Neural Network," Master's thesis, Queen Mary University of London (United Kingdom), 2020.

[3] A. Agudo, F. Moreno-Noguer, B. Calvo, and J. M. M. Montiel, "Sequential non-rigid structure from motion using physical priors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 979–994, 2015.

[4] D. Kim, M. Park, and Y.-L. Park, "Probabilistic modeling and Bayesian filtering for improved state estimation for soft robots," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1728–1741, 2021.

[5] A. El Amin, A. El-Rabbany, *et al.*, "Monocular VO scale ambiguity resolution using an ultra low-cost spike rangefinder," *Positioning*, vol. 11, no. 04, p. 45, 2020.

[6] M. Li, J. Yang, and L. Kneip, "Relative Pose for Nonrigid Multi-Perspective Cameras: The Static Case," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 96–105.

[7] M. Bosse, R. Zlot, and P. Flick, "Zebedee: Design of a spring-mounted 3-d range sensor with application to mobile mapping," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1104–1119, 2012.

[8] R. Zlot and M. Bosse, "Three-dimensional mobile mapping of caves." *Journal of Cave & Karst Studies*, vol. 76, no. 3, 2014.

[9] L. Martínez, J. Ruiz-del Solar, L. Sun, J. P. Siebert, and G. Aragon-Camarasa, "Continuous perception for deformable objects understanding," *Robotics and Autonomous Systems*, vol. 118, pp. 220–230, 2019.

[10] Y. Li, F. Chen, J. Cao, R. Zhao, X. Yang, X. Yang, and Y. Fan, "MLP Based Continuous Gait Recognition of a Powered Ankle Prosthesis with Serial Elastic Actuator," *arXiv preprint arXiv:2309.08323*, 2023.

[11] M. Lahariya, C. Innes, C. Develder, and S. Ramamoorthy, "Learning physics-informed simulation models for soft robotic manipulation: A case study with dielectric elastomer actuators," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 031–11 038.

[12] R. Tian, Y. Zhang, D. Zhu, S. Liang, S. Coleman, and D. Kerr, "Accurate and robust scale recovery for monocular visual odometry based on plane geometry," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5296–5302.

[13] M. Hutter, C. D. Remy, M. A. Hoepflinger, and R. Siegwart, "High compliant series elastic actuation for the robotic leg Scarl ETH," in *Field robotics*. World Scientific, 2012, pp. 507–514.

[14] T. Hinzmann, T. Taubner, and R. Siegwart, "Flexible stereo: constrained, non-rigid, wide-baseline stereo vision for fixed-wing aerial platforms," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2550–2557.

[15] T. Hinzmann, C. Cadena, J. Nieto, and R. Siegwart, "Flexible trinocular: Non-rigid multi-camera-IMU dense reconstruction for UAV navigation and mapping," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1137–1142.

[16] P. Foehn, E. Kaufmann, A. Romero, R. Penicka, S. Sun, L. Bauersfeld, T. Laengle, G. Cioffi, Y. Song, A. Loquercio, *et al.*, "Agilicious: Open-source and open-hardware agile quadrotor for vision-based flight," *Science robotics*, vol. 7, no. 67, p. eabl6259, 2022.

[17] J. K. Hopkins, B. W. Spranklin, and S. K. Gupta, "A survey of snake-inspired robot designs," *Bioinspiration & biomimetics*, vol. 4, no. 2, p. 021001, 2009.

[18] J. Liu, I. Dukes, and H. Hu, "Novel mechatronics design for a robotic fish," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2005, pp. 807–812.

[19] X. Peng, J. Cui, and L. Kneip, "Articulated multi-perspective cameras and their application to truck motion estimation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2052–2059.

[20] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[21] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[22] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE transactions on robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[23] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[25] L. Pan, D. Barath, M. Pollefeys, and J. L. Schönberger, "Global Structure-from-Motion Revisited," in *European Conference on Computer Vision (ECCV)*, 2024.

[26] J. L. Schönberger, T. Price, T. Sattler, J.-M. Frahm, and M. Pollefeys, "A Vote-and-Verify Strategy for Fast Spatial Verification in Image Retrieval," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[27] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[28] T. Schneider, M. T. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An Open Framework for Research in Visual-inertial Mapping and Localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1418–1425, 2018.

[29] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 5135–5142.

[30] T. Shan and B. Englot, "Lego-loam: Lightweight and ground-optimized lidar odometry and mapping on variable terrain," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 4758–4765.

[31] J. Zhang, S. Singh, *et al.*, "LOAM: Lidar odometry and mapping in real-time." in *Robotics: Science and systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[32] S. Cao, X. Lu, and S. Shen, "GVINS: Tightly coupled GNSS–visual–inertial fusion for smooth and consistent state estimation," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2004–2021, 2022.

[33] R. Murai, E. Dexheimer, and A. J. Davison, "MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16 695–16 705.

[34] D. Maggio, H. Lim, and L. Carlone, "Vggt-slam: Dense rgb slam optimized on the sl (4) manifold," *arXiv preprint arXiv:2505.12549*, 2025.

[35] G. Zheng, Y. Zhou, and M. Ju, "Robust control of a silicone soft robot using neural networks," *ISA transactions*, vol. 100, pp. 38–45, 2020.

[36] X. Wang, J. J. Dabrowski, J. Pinskier, L. Liow, V. Viswanathan, R. Scalzo, and D. Howard, "PINN-ray: A physics-informed neural network to model soft robotic fin ray fingers," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 247–254.

[37] M. I. Friswell, *Dynamics of rotating machines*. Cambridge university press, 2010.

[38] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM." https://github.com/MichaelGrupp/evo, 2017.