Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Smart capacitive e-skin takes soft robots beyond proprioception

Delin Hu
The University of Edinburgh
Francesco Giorgio-Serchi
The University of Edinburgh
Shiming Zhang
The University of Hong Kong
Yunjie Yang (y.yang@ed.ac.uk)
The University of Edinburgh

Research Article

Keywords: soft robotics, soft capacitive sensors, transformer, deformation reconstruction

Posted Date: April 6th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1501113/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Smart capacitive e-skin takes soft robots beyond proprioception

Delin Hu,[1] Francesco Giorgio-Serchi,[2,3] Shiming Zhang,[4]
Yunjie Yang[1*]

[1] SMART Group, Institute for Digital Communications, School of Engineering,
The University of Edinburgh, Edinburgh, UK
[2] Institute for Integrated Micro and Nano Systems, School of Engineering
The University of Edinburgh, Edinburgh, UK
[3] Edinburgh Centre for Robotics
[4] Department of Electrical and Electronic Engineering, The University of Hong Kong,
Hong Kong SAR, China

*To whom correspondence should be addressed; E-mail: y.yang@ed.ac.uk

Proprioception, the ability to perceive posture, position, and movement, is a prerequisite for accurate body control in animals and robots. The lack of three-dimensional full-body proprioceptive sense has become a key constraint for fundamental breakthroughs in bioinspired soft robots. Here, we report a type of flexible capacitive e-skins to endow soft robots with high-resolution, full-geometry bodily awareness beyond existing proprioception systems. We demonstrate that our intrinsically-stretchable e-skin can capture complex three-dimensional deformations across the entire soft body through recursive, multi-position capacitance measurements. The e-skin signals can be directly translated to high-density point clouds portraying the complete ge-

ometry via the deep net. This augmented proprioception system providing millimeter-scale, global sensory-motor awareness can offer a crucial solution to fundamental problems in soft robotics, such as enabling precise closed-loop control and developing digital twins of soft robots.

Introduction

The neuro-proprioceptive system of animals mediates the perception of the body's state and self-motion, constituting the prerequisite for precise and fast limbs coordination during locomotion and interaction with the environment (*1*). Similarly, the dexterous manipulation of intelligent robots relies on kinematic state estimation from the artificial proprioceptive system. Within the frame of conventional mechatronics, existing sensing technology already provides viable solutions to implementing artificial proprioceptive systems that meet the requirements of even the most agile and complex robotic platforms. This is due to the inherent predictability of the rigid-body system, whose finite number of degrees of freedom allow for full state estimation by means of a bounded set of distributed geometrical measurements. However, the development of artificial proprioceptive systems for highly deformable structures, such as soft robots, remains a fundamental challenge and severely restricts the observability and, ultimately, the capability to perform precise closed-loop control (*2*).

The elastic nature of soft robots represents their asset as well as their drawback. The bodily compliance of soft robots may provide an answer to the limits of conventional robots with respect to safety, adaptability and operational flexibility (*3, 4*), thus highlighting their spontaneous vocation for biomedical applications (*5–7*), human-robot interaction (*8–10*) as well as their employment in unstructured, potentially cluttered scenarios. However, this very feature also makes it infeasible to completely describe the 3D morphology of soft systems with only a few parameters, as thoroughly documented in (*3, 11*). To obtain comparable bodily awareness

with their rigid counterparts, soft robots require an augmented proprioceptive system that can capture geometric variations throughout the body and achieve full-geometry 3D shape reconstruction. Such augmented proprioception capability can provide more natural interaction with humans (e.g., real-time geometry of a soft robot can be visually observed without the line of sight) and underpin precise closed-loop control.

Thus far, soft robot proprioception is still in its infancy (*3*). Previous studies have focused on simplified proprioception problems, e.g., tracking a subset of topological entities or local geometrical features which offer a proximal representation of the body's state (*11–20*). Optical fibres embedded in the elastic medium of a slender structure, for example, successfully provide a direct description of the body's global curvature (*12*) and bending angle under bidirectional bending (*13*). In (*11*), combined readouts from multiple resistive sensors have shown to yield a reliable estimate of the 3D coordinates of a soft manipulator's end-effector. These methods provide partial descriptions of the 3D morphology of the soft body. Some recent studies have attempted to leverage machine learning to reconstruct certain simple deformations of soft robots under various constraints (*13, 17–19*). However, the state of the art is still far from a readiness level suitable for practical adoption.

Here we propose an augmented proprioception system to confer full-geometry, millimeter-level bodily awareness to soft robots which is beyond existing proprioception systems. The augmented proprioception system encapsulates an intrinsically stretchable capacitive e-skin (SCAS) and a purpose-designed neural architecture (i.e., the capacitance-to-deformation transformer, C2DT). The SCAS has four different functional layers (see Fig.1a) and employs a redundant planar skin electrode layout (see Fig.1a and b) that forms a sequence of capacitors sensitive to deformations across distal and proximal locations, allowing it to detect geometric variations across the entire soft body. The C2DT based on self-attention mechanism (*21*) explores the dependency over the e-skin signals and directly translates the measurements to the

full-geometry description of the morphology (see Fig.1c). The synergistic combination of the SCAS and C2DT can achieve accurate (mm-scale error) and high-resolution 3D deformation reconstruction under complex deformations. The proposed system is in principle agnostic to the shape of the soft body, as it requires no *a priori* geometrical knowledge, making it generalizable to soft robotics platforms with unprescribed morphology. This augmented proprioception capability can assist in solving the most fundamental challenges in soft robotics, such as precise closed-loop control in complex tasks, thereby facilitating their widespread adoption.

Results

Design of the e-skin in virtual environment We first design the SCAS in the virtual environment and quantify its performance, before its physical implementation (see Fig.S1). The virtual SCAS has a redundant layout of planar stretchable electrodes (the 64-electrode SCAS) to characterize the 3D deformation of the entire soft body (see Fig.S2). We implement 3D solid mechanics and electrostatics coupling field (3D-SECF) simulation to simultaneously model the e-skin response and soft body deformation. Considering the need to test the broadest range of possible deformations that are not achievable in a fully internally actuated system, we adopt a square cylindrical soft manipulator actuated by external forces as the testbed. Fig.S2A shows the geometric structure and electrode layout of the 3D-SECF model.

Any two SCAS electrodes can form a capacitor and the capacitance is sensitive to electrode deformations. The 64-electrode SCAS can theoretically produce 2,016 independent capacitance readouts in one measurement frame. We only record capacitances formed by electrodes in the same layer and those between two adjacent layers to ensure that they are practically measurable. Details of the measurement strategy are shown in Fig.S2B-S2C. Each SCAS measurement frame comprises 392 independent capacitance readouts. We argue that capacitances formed by these non-redundant combinations of SCAS electrodes contain sufficient information to portray

full-geometry deformations as their receptive fields cover the entire soft body.

Dynamic 3D-SECF simulation allows us to mimic a wide range of deformations and corresponding e-skin responses. We therefore generate a large-scale virtual proprioception dataset containing 39,334 samples. Each sample consists of a 3D point cloud with 1,716 points representing the deformation and corresponding 392 capacitance readouts. The deformations are driven by different external force loads, which can be divided into four different categories according to the types of external force loads, i.e., the combination of elongation and twisting $L_{(z,r)}$, pure bending $L_{(x,y)}$, two-phase twisting and bending $L_{r,(x,y)}$ and the combination of twisting and bending $L_{(x,y,r)}$. Simulation results show that SCAS signals can reflect the soft robot geometric variation under various complex deformations (see Fig.S3-S6), indicating its feasibility as proprioceptors. We then leverage the virtual proprioception dataset to quantify the SCAS performance in full-geometry 3D deformation reconstruction. The results are utilized to guide the optimal design of the physical SCAS and learning-based proprioception algorithms.

Capacitance-to-deformation transformer We employ 3D dense point clouds to represent the full-geometry morphology of the soft robot arm. We then consider deformation reconstruction as a set-to-set problem, mapping a SCAS signal set consisting of 392 capacitance readouts to its corresponding point set (a point cloud) in 3D space. Therefore, we propose a capacitance-to-deformation transformer (C2DT) based on self-attention mechanism (*21*) that is widespread in natural language processing (*22, 23*) and computer vision (*24, 25*) and shows superior performance in solving set-to-set problems. The framework of the C2DT is shown in Fig.2. C2DT infers the displacement of each point in the source point cloud (the one without deformation) from the proprioceptive information contained in SCAS signals. Given characteristics of electric field distribution, we hypothesize that capacitances from different electrode pairs convey different geometrical structure information. This is critical for the network to effectively distil

discriminative proprioceptive representations from capacitance readouts (*26*). We therefore design a special position encoding process in the C2DT to generate geometrical representations based on positions of individual electrode pairs (see Materials and Methods for more details).

We train a C2DT using the virtual proprioception dataset by minimizing the loss function consisting of the squared distance term of visual markers (for which point-to-point correspondences are known) and the Chamfer distance term of the remaining points (for which point-to-point correspondences are unknown); see Materials and Methods for details of the loss function, visual markers and training. The reconstruction results show superior quality (mm-scale error) and are able to capture the whole range of complex deformations tested (see Fig.3, Fig.S3-S6 and Movie S1). We employ four error metrics to quantitatively evaluate the performance of C2DT, i.e., the average distance (AD), the maximal distance (MD), the Chamfer distance (*27*) (CD) and the Hausdorff distance (*28*) (HD); see Materials and Methods for expressions of these metrics. We train several C2DTs with different hyperparameters and compare their performance to determine an optimized network structure. The quantitative results are shown in Table S1. We find that the C2DT with 6 backbone layers outperforms the other candidates. The AD error achieved with this setup is as low as 1.379 ± 1.048 mm, which is comparable to the accuracy achieved with RGB-D cameras frequently used as ground truth in the relevant research (*28*).

We implement ablation studies of the C2DT with 6 backbone layers to better understand the role of each loss term and position encoding. The results are shown in Fig.3, Fig.S3-S6 and Table S2. We observe that the C2DT cannot learn correct point-to-point correspondences without including visual markers in training. This phenomenon is illustrated in Fig.3, where the points in the region of interest of the source point cloud are not mapped into the correct corresponding region using the C2DT *w/o* markers. Although reconstructions show similarities with the

ground truth by minimizing the Chamfer distance term, point-to-point errors remain large. We also identify that by retaining only the squared distance term of the visual markers during training, local distortions arise in a set of frames of reconstructions. This indicates that the Chamfer distance term can benefit the geometrical quality of the reconstructions. Finally, we observe poor convergence when attempting to train the network after removing the position encoding part. We visualize the position representations of the trained C2DT through t-SNE (*29*) (Fig.S7) and see that after position encoding the electrode pairs with high geometrical correlation tend to cluster together, and the electrode pairs that are geometrically far apart are also far apart in the feature space. It suggests that our position encoder can generate distinctive geometric representations based on the locations of the input electrode pair.

The redundant SCAS design validates its feasibility in the virtual environment. However, the high density of markers and electrodes poses practical challenges to the fabrication and experimentation in the physical system. In order to reduce complexity while maintaining proprioception performance, we investigate the impact of the number of markers and electrode layout on the performance of the C2DT. The results of this analysis, shown in Fig.4A, prove the accuracy improvement from increasing the number of markers plateaus at 16. It provides evidence that a small set of markers is sufficient for the C2DT to establish correct point-to-point correspondences. Similarly, the reconstruction performance improves with the density of electrodes, but the improvement is minimal after the number of electrodes exceeds a certain value (e.g., 32), as illustrated in Fig.4B. These results highlight a favorable trade-off between reconstruction accuracy and electrode/markers units, confirming that it is safe to sacrifice a minute reduction in performance to simplify the fabrication and deployment of the SCAS.

Fabrication and characterization of the e-skin Based on the conclusions from the above investigation, we design a physical SCAS with 32 electrodes, consisting of 8 4-electrode SCAS modules. This design balances the full-geometry reconstruction performance with fabrication complexity. We fabricate multiple 4-electrode SCAS modules in parallel using established elastomer processing technologies (*30*). The electrodes are made of carbon black (CB) dispersed elastomers. However, this material is unsuitable for wires and interfaces due to its high resistance and non-linear, irreversible conductivity response under deformations (*31, 32*). Therefore, Eutectic Gallium 75.5% Indium 24.5% (EGaIn) is employed to fabricate the wires and interfaces due to its high conductivity (3.4×10^7 S m^{-1}) and stable response to deformation. The fabrication process is presented step by step in Fig.S8, and additional details regarding materials and fabrication are reported in Materials and Methods.

The 4-electrode SCAS module (20×20×120 mm) consists of 4 different functional layers (Fig.1A and 5A), i.e., the protective substrate (about 0.39 mm thick), the electrode layer (0.08 mm), the isolation layer (0.24 mm), and the sealing layer (0.3 mm). We engrave microchannels for wires (0.5 mm width) and connections (3×2 mm) on the isolation layer using a laser machine. Then the sealing layer is bonded to the outward surface of the isolation layer. We inject the EGaIn ink into the micro-channels. The CB electrodes and EGaIn wires are connected by vertical interconnect holes. The relative capacitance response of a 40% strain ranges from 16% to 19%, depending on the activated electrode pairs (the platform for cycling characterization is shown in Fig.S9). The response curves show excellent linearity and consistency over multiple cycles (more than 500 cycles in Fig.S10A-B). For comparison, we characterize a SCAS with CB wires using the same approach (see Fig.S11A-B). As Fig.5B illustrates, the SCAS with EGaIn wires is superior to its CB wires counterpart in terms of sensitivity (larger responses under the same deformations), linearity (no distortions in response curves) and cycling stability (does not

shift after 500 cycles of stretches).

We uniformly deploy 8 4-electrode SCAS modules on the surface of a soft manipulator with the size of 20×20×240 mm (see Fig.S12A and Fig.S13). The 32-electrode SCAS, consisting of 8 SCAS modules, connects to an in-house developed data acquisition system (*33*) to measure capacitance values. We use two oppositely placed RGB-D cameras (Azure Kinect) to capture real-time, ground-truth 3D deformations of the robot in the colour point cloud format from two complementary views, and then fuse them in a single coordinate system. We dye the sides of the robot arm white as its original transparency negatively impacts the quality of data collected by RGB-D cameras. Sixteen yellow visual markers are placed to encourage the network to learn correct point-to-point correspondences during training (see Fig.S12B). The experiment platform (see Fig.S14) can synchronously record capacitance and point cloud data at a frame rate of around 30 fps.

The reliability of the SCAS allows us to record capacitance readouts frames (each frame comprises 76 independent readouts) when the robot arm is subject to arbitrary external loading applied via the bottom holder over a long period (we intermittently collect about 1,220 s of deformation data during a 10 h experiment). To demonstrate the superiority of our approach, we implement a random sequence of complex deformations, including bending, elongation, twisting and their combinations (Fig.5C), during the experiment. In most frames, point clouds collected by cameras can not represent full-geometry 3D deformations due to missing points caused by inevitable visual occlusion. We fill points by α shape reconstruction (*34*) for the frames with minor missing point issues and directly filter the frames that have severe occlusion. Then we obtain a total of 31,380 frames of data (see Methods for details of data acquisition and preprocessing). A set of samples in this dataset is shown in Fig.S15.

Real-world augmented proprioception The challenge of real-world augmented proprioception is exacerbated by the relatively poor quality of point clouds (restricted by the accuracy of cameras, occlusion, light conditions), noise in the SCAS signals, and imperfect synchronization between different devices. In order to compensate for these added sources of inaccuracy, we enhance the C2DT framework by increasing the number of input frames (N_i adjacent frames of SCAS readouts) and introducing a regularization term in its loss function to limit the distance change between neighbouring points before and after deformation. We train several C2DTs with different input frame numbers using the filtered real-world dataset. The full-geometry reconstruction performance improves as the input frames number increases and achieves the minimum error at 3 adjacent input frames (Fig.6). This improvement indicates that increasing the number of input frames can reduce the negative impacts of noise in SCAS signals and asynchronization between devices. The temporal correlation among adjacent frames can also be considered favorable for deformation reconstruction. A representative set of reconstructions of the C2DT with 3 input frames is shown in Fig.7A-C and Movie S2. The results achieve 2.322 ± 0.687 mm for the CD metric (Table S3), comparable to the point clouds collected by RGB-D cameras. Compared to simulation results, some elaborate geometrical features of reconstructed deformations in certain frames are less obvious, especially for those related to twisting (Movie S3). This is mainly because the ground truth point clouds acquired by RGB-D cameras cannot reach the quality of point clouds synthesized by 3D-SECF simulation.

According to the ablation study (Fig.S16), visual markers play a similar role in physical and virtual environments, facilitating the network to learn correct point-to-point correspondences. We also observe that adding the neighbour regularization term can slightly improve the recon-

struction quality (Fig.S16). The position encoding part is crucial to extract useful proprioceptive information from physical SCAS signals. Similar to its contribution in the simulation, position encoding can assign discriminative high-dimensional representations to different electrode pairs based on their geometrical structures (Fig.8).

Discussion

We presented an augmented proprioception system that could visualize (i.e., present in 3D point clouds) 3D full-geometry deformations of soft robots. It is empowered by an intrinsically stretchable capacitive e-skin (SCAS) that leverages capacitances formed by the combinations of planar boundary electrodes, and an end-to-end neural architecture to translate SCAS signals directly into point clouds. While we demonstrated the advancement of this augmented proprioception approach on complex, continuous deformations, several issues remain to be addressed to further exploit its potential.

The SCAS fabrication involves manual operation (e.g., liquid metal injection, sealing layer attachment, interface to sensing electronics), leaving room for performance improvement. Although calibration of sensor readouts can, to a certain extent, mitigate this issue, a desirable solution would require automated manufacturing technologies, such as direct writing of liquid metal and 3D printing of soft materials.

While this work focuses on proprioception induced by actuation irrespective of external stimuli, real-world operation entails a spectrum of stimuli from the environment. These may, in turn, be the source of the deformations (e.g., compression) of the soft body as well as involve peripheral information (e.g., temperature, texture). Therefore, external stimuli perception is non-negligible. Due to its capacitive nature, the SCAS presented can, in principle, detect several types of external stimuli, such as tactile mapping and permittivity of the objects in the proximity

of the robot, offering a venue to combine the interpretation of endogenic and exogenic factors affecting the body's state. Furthermore, multi-modal sensor units can be integrated into the SCAS, which will enhance its capability to simultaneously detect multiple external stimuli.

We also point out that the C2DT belongs to the paradigm of supervised learning that requires abundant labelled data for training. A notorious problem is that the acquisition of labelled data is expensive, time-consuming and in some cases even impossible. For instance, point clouds of compression-induced deformation cannot be easily collected through vision-based methods due to inevitable occlusion. The proposed coupling field simulation can generate a large number of high-quality labeled training samples. However, the gap between virtual and physical environments leads to performance deterioration if the network is trained on the simulation dataset. This issue is common in many other fields. Sim-to-real transfer are considered as the potential solution. The development and application of sim-to-real approaches suitable for soft robot proprioception can significantly increase the value of the simulation data and reduce the cost of real-world data acquisition.

Notwithstanding the above limitations, the proposed augmented proprioception system can achieve real-time (30 fps), full-geometry deformation reconstruction with high accuracy (mm-scale error) and resolution (dense point clouds) under complex deformations. This level of proprioception represents a step change over existing attempts and is beyond existing proprioception systems. Notably, the system can be readily extended to different types of soft bodies through a straightforward learning process without requiring *a priori* knowledge. Implementing such high-resolution, full-geometry proprioception is essential for perceiving full-body status and achieving precise closed-loop control of soft robots, the key to breakthroughs in performing complex tasks.

Materials and Methods

Electrostatics-solid mechanics coupling field simulation The coupling field simulation is implemented in COMSOL Multiphysics to simultaneously generate virtual SCAS sensing data and deformation data to demonstrate the effectiveness of the proposed method. The object of study is a square soft robot arm made of silicone (length: 100 mm, width: 100 mm, height: 1000 mm, see Fig.S2A). An array of 64 electrodes (8×8) is placed on the surface of the robot arm to form a 64-electrode SCAS. For simplicity, each electrode is set as a 105×30 mm flat surface without thickness. The distance between two adjacent electrodes on the same side is 20 mm both horizontally and vertically. The distance between each edge and the nearest electrode is 10 mm. Relevant material properties are set as follows: Young's modulus $E = 4.15$ MPa, Poisson's ratio $v = 0.022$, density $\rho = 1.28{\times}10^3$ kg m^{-3}, relative permittivity $\varepsilon_r = 3$.

We implement 956 different episodes in the simulation to produce a virtual soft robot proprioception dataset. Each episode mimics a time-continuous deformation process and is discretized into about 40 frames. In each frame, the deformation and the corresponding capacitance readouts of the SCAS are recorded. We apply four different types of loads to generate various complex deformations: 1. *the combination of elongation and twisting* $L_{(z,r)}$: A torsion force and a pulling force along the z-axis are simultaneously applied to the tip of the robot arm; 2. *pure bending* $L_{(x,y)}$: A pulling force in the x-y plane is applied on the tip of the arm; 3. *two-phase twisting and bending* $L_{r,(x,y)}$: A torsion force is applied on the tip of the arm in the first r frames (r ranging from 6 to 16), and then a pulling force in the x-y plane is applied on the tip while maintaining the twisting state; 4. *the combination of twisting and bending* $L_{(x,y,r)}$: A torsion force and a pulling force in x-y plane are applied to the tip at the same time. Each deformation is represented by a 3D point cloud with 1,716 points. Examples are shown in Fig.S3-S6. Since it is impractical to ascertain the exact point-to-point correspondences of all

points between two different deformations in the real world, we resort to a scheme that can be realistically implemented in the physical experiment. We only select 64 points as virtual visual markers, whose correspondences are available during network training and the correspondences of the remaining points are only used in testing for evaluation (see Fig.S2A).

Theoretically, any two electrodes can form a capacitor. The SCAS with 64 electrodes can produce 2,016 independent capacitance readouts in each measurement frame. However, many of them are extremely small and cannot be reliably measured in the real world. Therefore, only capacitances of electrode pairs in the same layer and capacitances of certain electrode pairs between two adjacent layers are recorded. Fig.S2B shows all 28 electrode pairs in the first layer that form measurable independent capacitors. Fig.S2C shows all 24 electrode pairs between the first and second layers that form measurable independent capacitors. Following this sensing scheme, the SCAS can generate 392 independent capacitance readouts per measurement frame. Each readout is calibrated as follows: $c = (c' - c_{\text{emp}})/c_{\text{emp}}$, where c is the calibrated capacitance readout; c' is the original readout and c_{emp} is the readout without deformation. Examples of calibrated capacitance readouts are shown in Fig.S3-S6.

A total of 39,334 frames (956 episodes) of deformations and capacitance readouts are generated through the coupling field simulation, of which 2,319 frames (53 episodes) are with $L_{(z,r)}$; 12,552 frames (300 episodes) are with $L_{(x,y)}$; 12,269 frames (303 episodes) are with $L_{r,(x,y)}$ and 12,194 frames (300 episodes) are with $L_{(x,y,r)}$.

C2DT for the virtual SCAS In general, the C2DT is a deep model (Fig.2) that is able to deform the source point cloud $\boldsymbol{P}_s$ to approximate the target point cloud $\boldsymbol{P}$ based on the measurement characteristic tensor ($\boldsymbol{c}$, $\boldsymbol{Q}_{e_1}$, $\boldsymbol{Q}_{e_2}$). Here $\boldsymbol{P}_s \in \mathbb{R}^{N_p \times 3}$ is the point cloud without any deformation; N_p is the number of points in $\boldsymbol{P}_s$, which is 1,719 in this case; $\boldsymbol{P} \in \mathbb{R}^{N_p \times 3}$ and $\hat{\boldsymbol{P}} \in \mathbb{R}^{N_p \times 3}$ are the ground truth and reconstructed point clouds with a specific deformation

respectively; $\boldsymbol{c} \in \mathbb{R}^{N_m \times 1}$ is the corresponding calibrated capacitance readouts vector; N_m is the number of readouts in $\boldsymbol{c}$ with the value of 392 in this case; $\boldsymbol{Q}_{e_1} \in \mathbb{R}^{N_m \times 3}$ and $\boldsymbol{Q}_{e_2} \in \mathbb{R}^{N_m \times 3}$ are the coordinates of electrodes to generate $\boldsymbol{c}$.

The C2DT architecture consists of two parts, i.e., encoding and decoding. The input of the encoding part is $\boldsymbol{c}$, $\boldsymbol{Q}_{e_1}$ and $\boldsymbol{Q}_{e_2}$. $\boldsymbol{Q}_{e_1}$ and $\boldsymbol{Q}_{e_2}$ are considered as positional signals that can help distinguish different elements in $\boldsymbol{c}$. They pass through the multi-layer perceptron (MLP) $f_q(\cdot)$ to obtain the geometrical representations of individual electrodes. We next choose an element-wise max function to integrate the two electrode representations into the final geometrical representations for electrode pairs as the capacitance is independent of the order of electrodes according to the reciprocal theorem. The MLP $f_c(\cdot)$ maps $\boldsymbol{c}$ to high-dimensional representations, and the sum of capacitive and geometrical representations is the input of the transformer encoder $E(\cdot)$ with the length of N_m. For the decoding part, $\boldsymbol{P}_s$ is first fed to the MLP $f_s(\cdot)$, and then multi-head attention is implemented over the outputs of $f_s(\cdot)$ and $E(\cdot)$ through the transformer decoder $D(\cdot)$. The MLP $f_d(\cdot)$ is used to map the output sequence of $D(\cdot)$ to the displacement of each point, and the reconstruction $\hat{\boldsymbol{P}}$ is obtained by adding it to $\boldsymbol{P}_s$.

$\hat{\boldsymbol{P}}$ is expected to be as close as possible to the target point cloud $\boldsymbol{P}$. This goal is achieved by minimizing the following loss function:

$$\mathcal{L} = \mathbb{E}_{P \sim \mathcal{P}} \left[\lambda_1 \underbrace{\sum_{i=1}^{N_v} |\boldsymbol{p}_v^i - \hat{\boldsymbol{p}}_v^i|_2^2}_{\text{squared distance}} + \lambda_2 \underbrace{\sum_{j=1}^{N_r} \left(\min_{\boldsymbol{p}_r \in \boldsymbol{P}_r} |\boldsymbol{p}_r - \hat{\boldsymbol{p}}_r^j|_2^2 + \min_{\hat{\boldsymbol{p}}_r \in \hat{\boldsymbol{P}}_r} |\boldsymbol{p}_r^j - \hat{\boldsymbol{p}}_r|_2^2 \right)}_{\text{Chamfer distance}} \right] \tag{1}$$

where $\boldsymbol{P}_r \in \mathbb{R}^{N_r \times 3}$ represents the remaining points; $\boldsymbol{p}_r^j \in \mathbb{R}^3$ is the coordinates of the j^{th} remaining point; $\boldsymbol{p}_v^i \in \mathbb{R}^3$ is the coordinates of the i^{th} visual marker; N_v and N_r are the numbers of the visual markers and the remaining points respectively; $\mathcal{P}$ is the distribution of $\boldsymbol{P}$; λ_1 and λ_2 are the weights of the squared distance term of the visual markers and the Chamfer distance term of the remaining points, respectively.

The structures of subnetworks of the C2DT are as follows:

- f_s: Linear(3, h_{em})→ReLU → LayerNorm(h_{em}) → Linear(h_{em}, d_{model}) → ReLU → LayerNorm(d_{model})
- f_q: Linear(3, h_{em}) → ReLU → LayerNorm(h_{em}) → Linear(h_{em}, d_{model})
- f_c: Linear(1,h_{em}) → ReLU → LayerNorm(h_{em}) → Linear(h_{em}, d_{model})
- f_d: Linear(d_{model}, 3) → $a*$Tanh
- E: LayerNorm(d_{model}) → Transformer.EncoderLayer(d_{model}, d_{ff}, h, P_{drop})$\otimes n_{\text{e-layer}}$
- D: Transformer.MutualLayer(d_{model}, d_{ff}, h, P_{drop})$\otimes n_{\text{m-layer}}$ → Transformer.DecoderLayer(d_{model}, d_{ff}, h, P_{drop})$\otimes n_{\text{d-layer}}$

where h_{em}=32, d_{model}=128, a=1.2, d_{ff}=256, h=8, P_{drop}=0.1, $n_{\text{e-layer}}$=3, $n_{\text{m-layer}}$=1 and $n_{\text{d-layer}}$=2. Linear layers in f_q and f_c do not have learnable biases while others have. The LayerNorm in E takes the sum of capacitive and geometrical representations as input. Transformer.EncoderLayer and Transformer.DecoderLayer are exactly the same with the original transformer (*21*). We remove the first self-attention cell of Transformer.DecoderLayer and use the remaining part as Transformer.MutualLayer because $\boldsymbol{P}_s$ remains constant. Transformer.EncoderLayer$\otimes n_{\text{e-layer}}$ represents a stack of $n_{\text{e-layer}}$ Transformer.EncoderLayer.

We split the virtual proprioception dataset into three exclusive parts, i.e., training, validation and testing sets. The training set includes 22,517 frames (548 episodes), of which 1,334 frames (31 episodes) are with $L_{(z,r)}$, 7,204 frames (172 episodes) are with $L_{(x,y)}$, 6,980 frames (173 episodes) are with $L_{r,(x,y)}$ and 6,999 frames (172 episodes) are with $L_{(x,y,r)}$. The validation set includes 9,721 frames (236 episodes), of which 550 frames (12 episodes) are with $L_{(z,r)}$, 3,093 frames (74 episodes) are with $L_{(x,y)}$, 3,098 frames (76 episodes) are with $L_{r,(x,y)}$ and 2,980 frames (74 episodes) are with $L_{(x,y,r)}$. The test set includes 7,096 frames (172 episodes), of which 435 frames (10 episodes) are with $L_{(z,r)}$, 2,255 frames (54 episodes) are with $L_{(x,y)}$, 2,191 frames (54 episodes) are with $L_{r,(x,y)}$ and 2,215 frames (54 episodes) are with $L_{(x,y,r)}$.

The C2DT is implemented in PyTorch (*35*). We use the Adam (*36*) optimizer (β_1=0.9, β_2=0.98, ϵ=10^{-9}) to update learnable parameters and minimize $\mathcal{L}$. We set the initial learning rate of 0.001, which we decay by a factor of 1.2 every 15 epochs. We compute λ_1 and λ_2 as follows: $\lambda_1 = \lambda/3(\lambda N_v+2N_r)$, $\lambda_2 = 1/3(\lambda N_v+2N_r)$, where $\lambda = \max(1, 300-2*(\text{epoch}-1))$. We clip the gradient with the threshold of 0.5 and train the C2DT using the training set for 300 epochs with a batch size of 24. Each epoch takes about 9 min on 3 Nvidia Quadro P5000. We save the network with the least validation loss as our final model.

We quantitatively evaluate the performance of the C2DT through 4 error metrics, i.e., the average distance (AD), the maximal distance (MD), the Chamfer distance (CD) and the Hausdorff distance (HD):

$$\text{AD} = \frac{1}{N_p}\sum_{i=1}^{N_p}|\boldsymbol{p}^i - \hat{\boldsymbol{p}}^i|_2 \tag{2}$$

$$\text{MD} = \max_{i\in[1,N_p]}|\boldsymbol{p}^i - \hat{\boldsymbol{p}}^i|_2 \tag{3}$$

$$\text{CD} = \frac{1}{2N_p}\sum_{i=1}^{N_p}\left(\min_{\boldsymbol{p}\in\boldsymbol{P}}|\boldsymbol{p} - \hat{\boldsymbol{p}}^i|_2 + \min_{\hat{\boldsymbol{p}}\in\hat{\boldsymbol{P}}}|\boldsymbol{p}^i - \hat{\boldsymbol{p}}|_2\right) \tag{4}$$

$$\text{HD} = \max\left(\max_{\hat{\boldsymbol{p}}\in\hat{\boldsymbol{P}}}\min_{\boldsymbol{p}\in\boldsymbol{P}}|\boldsymbol{p} - \hat{\boldsymbol{p}}|_2, \max_{\boldsymbol{p}\in\boldsymbol{P}}\min_{\hat{\boldsymbol{p}}\in\hat{\boldsymbol{P}}}|\boldsymbol{p} - \hat{\boldsymbol{p}}|_2\right) \tag{5}$$

We compare the performance of C2DTs with different hyperparameters and the results are shown in Table S1. To understand the impact of each loss term and position encoding on the performance, we also implement ablation studies. We remove the squared distance term and the Chamfer distance term respectively and perform the same training procedure to obtain results of the C2DT *w/o* markers and the C2DT *w/o* Chamfer distance. The reconstructed point clouds and values of these metrics are shown in Fig.3, Fig.S3-S6 and Table S2. We also try to train the network without the position encoding part, but it is unable to converge. The position

representations of the trained C2DT is visualized through t-SNE (*29*) and presented in Fig.S7, which can help discover the geometrical correlation among different electrode pairs.

We also investigate the performance of C2DTs with different numbers of visual markers and different electrode layouts using the same method to guide the sensor and network design in the real world. The results are shown in Fig.4.

SCAS fabrication, characterization and deployment The 32-electrode SCAS comprises 8 modular 4-electrode SCASs. Each SCAS module has 4 different functional layers, i.e., the protective substrate, the electrode layer, the isolation layer and the sealing layer. We fabricate each SCAS module layer by layer. The steps are shown in Fig.S8: i) We mix Smooth-on Ecoflex 00-30 part A (1.0) and part B (1.0) and pour it on a glass plate. Then we use a TQC Sheen micrometer film applicator to flatten the silicone and cure it for 3 min at 100°C. ii) We first mix Imerys Enasco 250P conductive carbon black (0.2) with isopropyl alcohol (2.0), after which the uncured silicone mixture (2.0) is added and we stir them for 3 min. A layer of uncured conductive silicone is coated on the protective substrate and is cured for 3 min in a 100°C oven. iii) We use a 40 W Aeon MIRA 5 laser machine to pattern CB electrodes. The parameters are set as follows: 28% Power, 300 mm s^{-1} Speed and 0.05 mm Interval. The planar size of each electrode is 21 $\times$ 6 mm, which is one-fifth of the one studied in the simulation. iv) We use the same method as in step i to fabricate a silicone membrane for the isolation layer on the top of the electrode layer. v) Two rounds of engraving are performed with 20.5% Power, 300 mm s^{-1} Speed and 0.05 mm Interval to generate micro channels of liquid metal wires and connections to readout electronics. Four rounds of engraving are conducted with the same parameters to generate vertical interconnect holes. The planar size of readout connections and vertical interconnect holes is 3 $\times$ 2 mm, and the width of wires is 0.5 mm. Then we cut the rectangular area of the modular SCAS with 19.5% Power and 25 mm s^{-1} Speed and remove the

remaining part. vi) We fabricate a new silicone membrane following step i, and we uniformly coat a very thin layer of uncured silicone mixture on its surface as adhesive. Then we bond the SCAS cut in step v with the membrane. The curing takes about 4 h under room temperature to ensure high-quality bonding. vii) We inject Eutectic Gallium 75.5% Indium 24.5% (EGaIn, Sigma Aldrich) ink from readout connections, and meanwhile exhaust the air in microchannels through the vertical interconnect holes. viii) We obtain the final modular 4-electrode SCAS. The planar size of the SCAS module is 120 $\times$ 20 mm, of which 100 $\times$ 20 mm is the area of the electrodes, and 20 $\times$ 20 mm is the interface to readout electronics. The layer thicknesses are 0.39 mm, 0.08 mm, 0.24 mm and 0.3 mm, respectively. Since the fabrication is easy to scale up, we manufacture 5 SCAS modules in parallel.

To characterize the response of the SCAS module and verify the superior performance of EGaIn wires compared with CB wires, we attach a 4-electrode SCAS with CB wires and a 4-electrode SCAS with EGaIn wires on the front and back sides of a segment of the square cylinder silicone structure (20$\times$20$\times$140 mm) and cyclically stretch them using a Nema23 stepper motor with a SFU1605 ball screw (see Fig.S9). Each cycle takes 20 s, and the SCASs are strained by up to 40%. The entire test takes about 3 h (more than 500 cycles). Relative capacitance readouts of each SCAS are illustrated in Fig.S10-S11. The results show that the SCAS with EGaIn wires has better sensitivity (larger response under the same deformation), linearity (no distortions in response curves) and cycling stability (no drift after 500 cycles).

We cast a square cylinder robot arm with the size of 20 $\times$ 20 $\times$ 240 mm which is one-fifth of the one in the simulation. The extra 40 mm in height is the interface area for driving the deformation, connecting to electronics and bonding with the fixed ceiling. We bond 8 4-electrode SCAS modules on its surface to form the 32-electrode SCAS (see Fig.S12A). The transparency of silicone adversely impacts on the quality of the point clouds collected by RGB-D cameras based on the time-of-flight principle. We therefore coat a silicone layer with white

Smooth-on Silc Pig Silicone Pigments for better reflection. We also attach 16 yellow dots as visual markers to assist network training with correspondence information. We cover the interface to readout electronics with black acrylic tape to reduce its interference in point cloud collection (see Fig.S12B). Individuals electrodes are indexed and accessible from the readout electronics (Fig.S13).

Experimental setup The experiment platform consists of the soft robot arm equipped with the 32-electrode SCAS, the readout electronics, two Microsoft Azure Kinect RGB-D cameras (*37*) and a laptop installed with a customized software to control the readout electronics and record data from the cameras and the SCAS (see Fig.S14). The readout electronics is based on a 32-electrode ECT system that supports arbitrary switching schemes. Its capacitance measurement resolution is 3 fF, and the signal to noise ratio of all 32 channels is above 60 dB.

The two cameras are placed directly opposite and in a straight line with the robot arm to capture its 3D deformations from two complementary views in real time. The deformations are saved and represented via the colour point cloud format. The data recording of the cameras and readout electronics is synchronized. The frame rate can reach about 30 fps if we only record the point cloud and capacitance data. It will decrease to around 20 fps if RGB images are also recorded.

Experimental data acquisition and preprocessing In real-world experiments, we manipulate manually the hand holder bonding to the bottom of the robot arm to induce a variety of complex deformations, including elongation, twisting, omni-directional bending and their combinations (see Fig.5C, Fig.S15B-C). Meanwhile, we synchronously record the SCAS and cameras data (i.e., capacitance readouts, colour point clouds, and sometimes RGB images). We collect 36,465 frames (about 1,220 s) of experimental data. In this real-world dataset, the first 36,013 frames (about 1,200 s) record only the capacitance readouts and colour point clouds; the

last 452 frames (about 20s) also save the RGB images with a reduced frame rate.

The 32-electrode SCAS can produce 76 capacitance readouts in a single frame, which are calibrated using the same method as in the simulation. The point clouds from the two cameras are fused in one coordinate system using the chessboard calibration method (*38, 39*). The raw data is noisy and contains many meaningless background points, making it unusable for direct training. We clean and preprocess the data using Matlab (R2021a) to selectively retain only the points on the surface of the robot arm. The points on the black acrylic tape and red holders are eliminated via colour-filtering. In order to further reduce the negative impact of noise and outliers, we filter out regions whose local point densities are lower than a preset threshold. Due to inevitable visual occlusion, in many frames the cleaned point clouds cannot completely represent 3D deformations. To alleviate this issue, further preprocessing is required prior to training. We implement average grid downsampling with a 4 mm box gird filter at first for computational efficiency. Then we reconstruct α shapes (*34*) on the basis of the downsampled point clouds to alleviate the issue of incomplete representation. The triangular meshes of the alpha shapes are subdivided three times, and vertexes are extracted as new point clouds with supplementary points. In our C2DT framework, the numbers of points in the source and target point clouds are expected to be the same. In order to meet this requirement, we first implement average grid downsampling with a 4 mm box gird filter and then use farthest point sampling (*40*) to eventually select 1,300 points in each point cloud.

We extract yellow visual markers from cleaned point clouds before downsampling and α shape reconstruction based on the RGB information of each point. We create a graph according to one frame of marker points. The connection of each two points in the graph is determined by their distance. The threshold of connected distance is 6 mm. Each connected subgraph with more than 10 points is considered as a visual marker, and the average of the coordinates of all points in a subgraph is used to represent the marker position. The number of extracted visual

markers is not always 16 due to camera occlusion. It is almost impossible to automatically obtain point-to-point correspondences of visual markers under our current experimental setup. We therefore align visual markers layer-to-layer. The 16 visual markers can be divided into 4 layers, and each layer includes 4 markers. We create a graph based on one frame of coordinates of extracted markers with the connected distance threshold of 26 mm. Each subgraph is a layer of markers. The permutation of the layer is determined by the relative position in the y-axis of the fused coordinate system among all 4 layers. We delete all abnormal frames for which the number of extracted markers is larger than 16 and/or the number of layers is not equal to 4. We fill the layers for which the number of markers is less than 4 with (0,0,0) to ensure all layers have the same number of points, which can improve the computational efficiency during training. Furthermore, we remove the frames with critically missing points issues because of the low quality of their reconstructed α shapes. The number of markers in individual layers indicates the severity of missing points. The frames with at least 2 markers in all layers are retained while others are dismissed.

Upon the above filtering process, a total of 31,380 frames of data remains available for analysis. We randomly inspect a sample of 500 frames out of the dataset and do not find serious missing points issues.

C2DT for the physical SCAS The basic framework of C2DT in the real-world experiment is analogous to that in the simulation. However, some modifications are required due to the difference between the real and virtual environments. First, the loss function in simulation is no longer applicable, as in our experiments the point-to-point correspondences of visual markers

are not available. Instead, we propose a modified loss function as follows.

$$\mathcal{L}^{\star} = \mathbb{E}_{P\sim\mathcal{P}}\Bigg\{ \lambda_1 \sum_{k=1}^{N_l}\sum_{i=1}^{N_{lv}} \Big[d(\hat{\boldsymbol{p}}^i_{l_k}, \boldsymbol{P}_{l_k}) \cdot S^{k,i}_{r2g} + d(\boldsymbol{p}^i_{l_k}, \hat{\boldsymbol{P}}_{l_k}) \cdot S^{k,i}_{g2r} \Big] + \lambda_2 \sum_{j=1}^{N_r} \Big[d(\hat{\boldsymbol{p}}^j_r, \boldsymbol{P}_r) + d(\boldsymbol{p}^j_r, \hat{\boldsymbol{P}}_r) \Big] \\ + \lambda_3 \sum_{j=1}^{N_r}\sum_{l=1}^{N_n} \Big[\big(|\hat{\boldsymbol{p}}^j_r - \hat{\boldsymbol{p}}^{j,l}_r|_2 - \delta_d \cdot s^{j,l} \big)^2 \cdot S^{j,l}_d + \big(|\hat{\boldsymbol{p}}^j_r - \hat{\boldsymbol{p}}^{j,l}_r|_2 - \delta_u \cdot s^{j,l} \big)^2 \cdot S^{j,l}_u \Big] \Bigg\} \quad (6)$$

The first term of $\mathcal{L}^{\star}$ counts the Chamfer distance between the reconstruction and the ground truth of markers layer-by-layer, where $\boldsymbol{P}_{l_k} \in \mathbb{R}^{N_{lv}\times 3}$ is the coordinates of the visual markers in the l_k layer; $\boldsymbol{p}^i_{l_k} \in \mathbb{R}^3$ is the coordinates of the i^{th} point in $\boldsymbol{P}_{l_k}$; $d(\hat{\boldsymbol{p}}^i_{l_k}, \boldsymbol{P}_{l_k})$ is the squared distance between $\hat{\boldsymbol{p}}^i_{l_k}$ and the nearest point in $\boldsymbol{P}_{l_k}$; N_l is the number of layers; N_{lv} is the number of marker in each layer and the values of N_l and N_{lv} are 4 in this case. When computing the loss, we only need to consider the marker points extracted in the data preprocessing and ignore the padding points. Note that all points in $\hat{\boldsymbol{P}}_{l_k}$ are marker points as they are generated by the network based on the corresponding capacitance readouts and the source point, which does not include padding points. In order to eliminate the effect of padding points during training, we synthesize masks $S^{k,i}_{r2g}$ and $S^{k,i}_{r2g}$ as follows.

- $S^{k,i}_{g2r}$ is set to 1 if $\boldsymbol{p}^i_{l_k}$ is a marker point. $S^{k,i}_{g2r}$ is set to 0 if $\boldsymbol{p}^i_{l_k}$ is a padding point.
- $S^{k,i}_{r2g}$ is set to 1 if $\boldsymbol{P}_{l_k}$ does not include any padding points, otherwise $S^{k,i}_{r2g}$ is set to 0.

The second term in $\mathcal{L}^{\star}$ is exactly the same as its simulation counterpart that counts the Chamfer distance between the reconstruction and ground truth of the remaining points. The third term is a regularizer that encourages the distance between neighbouring points to not change significantly before and after deformations, where $\hat{\boldsymbol{p}}^{j,l}_r$ is the l^{th} neighbour of $\hat{\boldsymbol{p}}^j_r$; $s^{j,l}$ is the distance between the corresponding two points in the source point cloud; δ_d and δ_u are coefficients of thresholds. We count the loss only if the neighbour distance in the reconstruction falls outside the preset range. We achieve this with masks $S^{j,l}_d$ and $S^{j,l}_u$ as follows.

- $S_d^{j,l}$ is set to 1 if $|\hat{\boldsymbol{p}}_r^j - \hat{\boldsymbol{p}}_r^{j,l}|_2 - \delta_d \cdot s^{j,l} < 0$, otherwise $S_d^{j,l}$ is set to 0.

- $S_u^{j,l}$ is set to 1 if $|\hat{\boldsymbol{p}}_r^j - \hat{\boldsymbol{p}}_r^{j,l}|_2 - \delta_u \cdot s^{j,l} > 0$, otherwise $S_u^{j,l}$ is set to 0.

The number of input frames in the physical world is not constant to 1. In contrast, the C2DT takes several (N_i) adjacent frames as its input. The first linear cell in f_c is therefore modified to Linear(N_i,h_{em}). The hyper-parameters of the C2DT are set as: h_{em}=32, d_{model}=64, d_{ff}=128, h=4, P_{drop}=0.1, $n_{\text{e-layer}}$=2, $n_{\text{m-layer}}$=1, $n_{\text{d-layer}}$=1 and a=1.9472. We magnify the values of point cloud data five times to make it close to the simulation scale. The network is trained and evaluated using almost the same procedure as presented earlier.

We split the real-world dataset into three exclusive parts. The first 26,771 frames (about 1,020 s) are used for training (20,693 frames) and validation (6,018 frames), and the last 4,669 frames (about 200 s) are used for testing. We set δ_d=0.5 and δ_u=2. We compute λ_1, λ_2 and λ_3 as follows: $\lambda_1 = \lambda/[\lambda \sum_{k=1}^{N_l} \sum_{i=1}^{N_{lv}} (S_{r2g}^{k,i} + S_{g2r}^{k,i}) + 2N_r]$, $\lambda_2 = 1/[\lambda \sum_{k=1}^{N_l} \sum_{i=1}^{N_{lv}} (S_{r2g}^{k,i} + S_{g2r}^{k,i}) + 2N_r]$, $\lambda_3 = 1/10[\sum_{j=1}^{N_r} \sum_{l=1}^{N_n} (S_d^{k,i} + S_u^{k,i})]$, where $\lambda = \max(1, 300 - 10 * (\text{epoch} - 1))$. In total, we run 200 epochs with a batch size of 39 and retain the network with the least validation loss. We implement ablation studies to evaluate the effect of individual loss terms (see Fig.S14) and quantitatively evaluate reconstructions on the real-world scale with CD and HD metrics, which do not require point-to-point correspondences (Table S3). Finally, we visualize the position representations of individual electrode pairs via t-SNE to illustrate the geometrical correlation between different capacitance readouts (Fig.8).

Supplementary Materials and Methods

Tab S1: Quantitative evaluation of C2DTs with different network hyperparameters on the simulation dataset.

Tab S2: Quantitative evaluation of C2DTs with different loss terms on the simulation dataset.

Tab S3: Quantitative evaluation of C2DTs with different network hyperparameters on the real-world dataset.

Fig S1: Flowchart of coupling field simulation and 3D deformation reconstruction using a virtual 64-electrode SCAS.

Fig S2: Electrode layout and measurement strategy of the 64-electrode SCAS in the simulation.

Fig S3: A set of examples in one testing episode with a continuous $L_{(z,r)}$ deformation process.

Fig S4: A set of examples in one testing episode with a continuous $L_{(x,y)}$ deformation process.

Fig S5: A set of examples in one testing episode with a continuous $L_{r,(x,y)}$ deformation process.

Fig S6: A set of examples in one testing episode with a continuous $L_{(x,y,r)}$ deformation process.

Fig S7: Visualization of position representations in the simulation.

Fig S8: Fabrication process of a modular 4-electrode SCAS.

Fig S9: Cycling test platform.

Fig S10: Characterization of the 4-electrode SCAS with EGaIn wires.

Fig S11: Characterization of the 4-electrode SCAS with CB wires.

Fig S12: Deployment of the 4-electrode SCAS modules.

Fig S13: Schematic diagram of the 32-electrode SCAS that consists of 8 4-electrode SCAS modules.

Fig S14: Real-world experiment platform.

Fig S15: Samples in the real-world dataset.

Fig S16: Qualitative analysis of ablation studies results.

References

1 J. C.Tuthill and E. Azim, "Proprioception," *Current Biology*, vol. 28, no. 5, pp. 194–203, 2018.

2 B. Shih, D. Shah, J. Li, T. G. Thuruthel, Y.-L. Park, F. Iida, Z. Bao, R. Kramer-Bottiglio, and M. T. Tolley, "Electronic skins and machine learning for intelligent soft robots," *Science Robotics*, vol. 5, p. eaaz9239, 04 2020.

3 H. Wang, M. Totaro, and L. Beccai, "Toward perceptive soft robots: Progress and challenges," *Advanced Science*, 2018.

4 D. Rus and M. T. Tolley, "Design, fabrication and control of soft robots," *Nature*, vol. 521, pp. 467–475, 05 2015.

5 M. Cianchetti, C. Laschi, A. Menciassi, and P. Dario, "Biomedical applications of soft robotics," *Nature Reviews Materials*, vol. 3, no. 6, pp. 143–153, 2018.

6 N. Cheng, J. Amend, T. Farrell, D. Latour, C. Martinez, J. Johansson, A. McNicoll, M. Wartenberg, S. Naseef, W. Hanson, and W. Culley, "Prosthetic jamming terminal device: A case study of untethered soft robotics," *Soft Robotics*, vol. 3, pp. 205–212, 2021/07/11 2016.

7 Y.-L. Park, B. rong Chen, N. O. Pérez-Arancibia, D. Young, L. Stirling, R. J. Wood, E. C. Goldfield, and R. Nagpal, "Design and control of a bio-inspired soft wearable robotic device for ankle–foot rehabilitation," *Bioinspiration & Biomimetics*, vol. 9, p. 016007, jan 2014.

8 T. Arnold and M. Scheutz, "The tactile ethics of soft robotics: Designing wisely for human–robot interaction," *Soft Robotics*, vol. 4, pp. 81–87, 2021/07/11 2017.

9 A. Moin, A. Zhou, A. Rahimi, A. Menon, S. Benatti, G. Alexandrov, S. Tamakloe, J. Ting, N. Yamamoto, Y. Khan, F. Burghardt, L. Benini, A. C. Arias, and J. M. Rabaey, "A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition," *Nature Electronics*, vol. 4, no. 1, pp. 54–63, 2021.

10 X. Yu, Z. Xie, Y. Yu, J. Lee, A. Vazquez-Guardado, H. Luan, J. Ruban, X. Ning, A. Akhtar, D. Li, B. Ji, Y. Liu, R. Sun, J. Cao, Q. Huo, Y. Zhong, C. Lee, S. Kim, P. Gutruf, C. Zhang, Y. Xue, Q. Guo, A. Chempakasseril, P. Tian, W. Lu, J. Jeong, Y. Yu, J. Cornman, C. Tan, B. Kim, K. Lee, X. Feng, Y. Huang, and J. A. Rogers, "Skin-integrated wireless haptic interfaces for virtual and augmented reality," *Nature*, vol. 575, no. 7783, pp. 473–479, 2019.

11 T. G. Thuruthel, B. Shih, C. Laschi, and M. T. Tolley, "Soft robot perception using embedded soft sensors and recurrent neural networks," *Science Robotics*, vol. 4, no. 26, p. eaav1488, 2019.

12 C. To, T. L. Hellebrekers, and Y. Park, "Highly stretchable optical sensors for pressure, strain, and curvature measurement," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5898–5903, 2015.

13 I. M. Van Meerbeek, C. M. De Sa, and R. F. Shepherd, "Soft optoelectronic sensory foams with proprioception," *Science Robotics*, vol. 3, no. 24, p. eaau2489, 2018.

14 T. Kim, S. Lee, T. Hong, G. Shin, T. Kim, and Y.-L. Park, "Heterogeneous sensing in a multifunctional soft sensor for human-robot interfaces," *Science Robotics*, vol. 5, no. 49, 2020.

15 Y. Zhao, C.-Y. Lo, L. Ruan, C.-H. Pi, C. Kim, Y. Alsaid, I. Frenkel, R. Rico, T.-C. Tsao, and X. He, "Somatosensory actuator based on stretchable conductive photothermally responsive hydrogel," *Science Robotics*, vol. 6, no. 53, 2021.

16 R. B. N. Scharff, G. Fang, Y. Tian, J. Wu, J. M. P. Geraedts, and C. C. L. Wang, "Sensing and reconstruction of 3d deformation on pneumatic soft robots," *IEEE/ASME Transactions on Mechatronics*, 2021.

17 R. L. Truby, C. D. Santina, and D. Rus, "Distributed proprioception of 3d configuration in soft, sensorized robots via deep learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3299–3306, 2020.

18 O. Glauser, D. Panozzo, O. Hilliges, and O. Sorkine-hornung, "Deformation capture via soft and stretchable sensor arrays," *ACM Transactions on Graphics*, vol. 38, no. 2, 2019.

19 A. Leber, C. Dong, R. Chandran, T. Das Gupta, N. Bartolomei, and F. Sorin, "Soft and stretchable liquid metal transmission lines as distributed probes of multimodal deformations," *Nature Electronics*, vol. 3, no. 6, pp. 316–326, 2020.

20 C. Armanini, M. Farman, M. Calisti, F. Giorgio-Serchi, C. Stefanini, and F. Renda, "Flagellate underwater robotics at macroscale: Design, modeling, and characterization," *IEEE Transactions on Robotics*, pp. 1–17, 2021.

21 A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *the International Conference on Neural Information Processing Systems*, 2017.

22 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *roceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

23 Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

24 H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

25 L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.

26 D. Hu, K. Lu, and Y. Yang, "Image reconstruction for electrical impedance tomography based on spatial invariant feature maps and convolutional neural network," in *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, 2019.

27 H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

28 R. Wang, S. Wang, S. Du, E. Xiao, W. Yuan, and C. Feng, "Real-time soft body 3d proprioception via deep vision-based sensing," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3382–3389, 2020.

29 L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

30 O. A. Araromi, S. Rosset, and H. R. Shea, "High-resolution, large-area fabrication of compliant electrodes via laser ablation for robust, stretchable dielectric elastomer actuators and sensors," *ACS Applied Materials & Interfaces*, vol. 7, pp. 18046–18053, 08 2015.

31 S. H. Yoon, L. Paredes, K. Huo, and K. Ramani, "Multisoft: Soft sensor enabling real-time multimodal sensing with contact localization and deformation classification," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, Sept. 2018.

32 Z. Zhu, H. S. Park, and M. C. McAlpine, "3d printed deformable sensors," *Science Advances*, vol. 6, p. eaba5575, 06 2020.

33 Y. Yang, L. Peng, and J. Jia, "A novel multi-electrode sensing strategy for electrical capacitance tomography with ultra-low dynamic range," *Flow Measurement and Instrumentation*, vol. 53, pp. 67–79, 2017.

34 H. Edelsbrunner, D. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transactions on Information Theory*, vol. 29, no. 4, pp. 551–559, 1983.

35 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

36 D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

37 M. Tölgyessy, M. Dekan, Ľ. Chovanec, and P. Hubinský, "Evaluation of the azure kinect and its comparison to kinect v1 and kinect v2," *Sensors*, vol. 21, no. 2, 2021.

38 A. Geiger, F. Moosmann, Ö. Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *2012 IEEE International Conference on Robotics and Automation*, pp. 3936–3943, 2012.

39 V. Soleimani, M. Mirmehdi, D. Damen, S. Hannuna, and M. Camplani, "3d data acquisition and registration using two opposing kinects," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 128–137, 2016.

40 C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), pp. 5105–5114, Curran Associates Inc., 2017.

Acknowledgements: Y.Y. and F.G. acknowledge the support of the Data Driven Innovation Chancellor's Fellowship at The University of Edinburgh. D.H. acknowledge the support of the studentship from the School of Engineering, The University of Edinburgh. S. Z. acknowledges the support of the start-up funds supported by The University of Hong Kong.

Author contributions: Y.Y., F.G., and D.H. conceived the concept. Y.Y. and F.G. supervised the project. D.H. and Y.Y. carried out the simulation, the fabrication and experiments. S.Z. guided the material fabrication and characterization. Y.Y. designed the measurement electronics and software. D.H. designed the machine learning algorithm. D.H. and Y.Y. analysed the data. D.H. wrote the manuscript. Y.Y., F.G. and S.Z. reviewed and edited the manuscript.

Competing interests: Y.Y., F.G., and D.H. are applying for patents related to the described work. The other authors declare that they have no competing interests.

Data availability: All data are available from the corresponding author upon reasonable request.

Code availability: Custom code used in the current study is available from the corresponding author on request.

Correspondence and requests for materials should be addressed to Y.Y.

Fig. 1. Design of the SCAS and the pipeline for full-geometry, high-resolution 3D deformation reconstruction of soft robots. (A) The entire SCAS that can cover the whole soft robot arm consists of multiple SCAS modules. Each module has 4 functional layers, i.e., the protective substrate (0.39 mm), the electrode layer (0.08 mm), the isolation layer (0.24 mm) and the sealing layer (0.3 mm). The soft electrodes are made of carbon black (CB) dispersed elastomers. Eutectic Gallium 75.5% Indium 24.5% (EGaIn) is employed to fabricate the wires and interfaces. The multi-position skin electrode combinations can form a sequence of capacitors. Geometric variations in the proximity of the electrode pair lead to the change of the corresponding capacitance. The readout electronics can record capacitance values of a selected set of electrode pairs at approximately 30 fps. (B) Snapshots of the soft arm in different states (undeformed, twisting and the combination of bending and twisting). (C) The data collected by the readout electronics is fed into a deep net and translated to a high-resolution representation of the full-body 3D shape of the soft arm.

Fig. 2. The architecture of the C2DT that infers the displacement of each point in the source point cloud (without any deformations) from the SCAS capacitance readouts. In the encoding part, the network encodes the input capacitance readouts and the geometrical structure information of electrode pairs to a high-dimensional space and feeds them to the transformer encoder to distil proprioceptive information. In the decoding part, the network manages to assign a correct displacement to each point in the source point cloud based on the output sequence of the encoding part. See Methods for more implementation and architecture details.

Fig. 3. A set of examples of reconstructions generated by different C2DTs. The colour of each point in reconstructions indicates the distance from the corresponding point in the ground truth. The region of interest is the middle section in the source point cloud. The points in the region of interest (marked in black) should be mapped into the middle section in reconstructions if C2DTs learn correct point-to-point correspondences. We can observe apparent shifts in the reconstructions of the C2DT *w/o* markers.

Fig. 4. Reconstruction performance under different marker numbers and different electrode layouts. (A)The performance of the C2DT under 4 different numbers of markers (mean$\pm$standard deviation). (B) The performance of the C2DT under 4 different electrode layouts (mean$\pm$standard deviation).

Fig. 5. Characterization of the SCAS. (A) 4-electrode SCAS modules with EGaIn (centre) and CB (right) and wires. Left, the cross-section of the module with EGaIn wires under a 40x digital microscope. (B) Relative capacitance response curves of the two SCAS modules to a 40% periodic linear stretch. The SCAS with EGaIn wires shows better sensitivity, linearity and cycling stability than its CB wires counterpart. (C) A representative set of complex multimodal deformations that appear in our experiment.

Fig. 6. The performance of C2DTs, which take different numbers of adjacent frames as inputs, is shown (mean$\pm$standard deviation).

Fig. 7. 3D full-geometry deformation reconstruction based on real-world SCAS signals. (A) Three representative frames of a bending process are shown in RGB images, ground truth point clouds collected by depth cameras and reconstructed point clouds of the C2DT. The colour of each point in reconstructions indicates the distance from the nearest point in the ground truth. Examples of other deformations

are shown in (B) the combination of elongation and twisting and (C) the combination of bending and twisting.

Fig. 8. Two-dimensional t-SNE embedding of the output of the position encoding part in the trained C2DT. The electrode pairs between two adjacent layers and the electrode pairs in the same layer are marked in red and blue respectively.

Figures

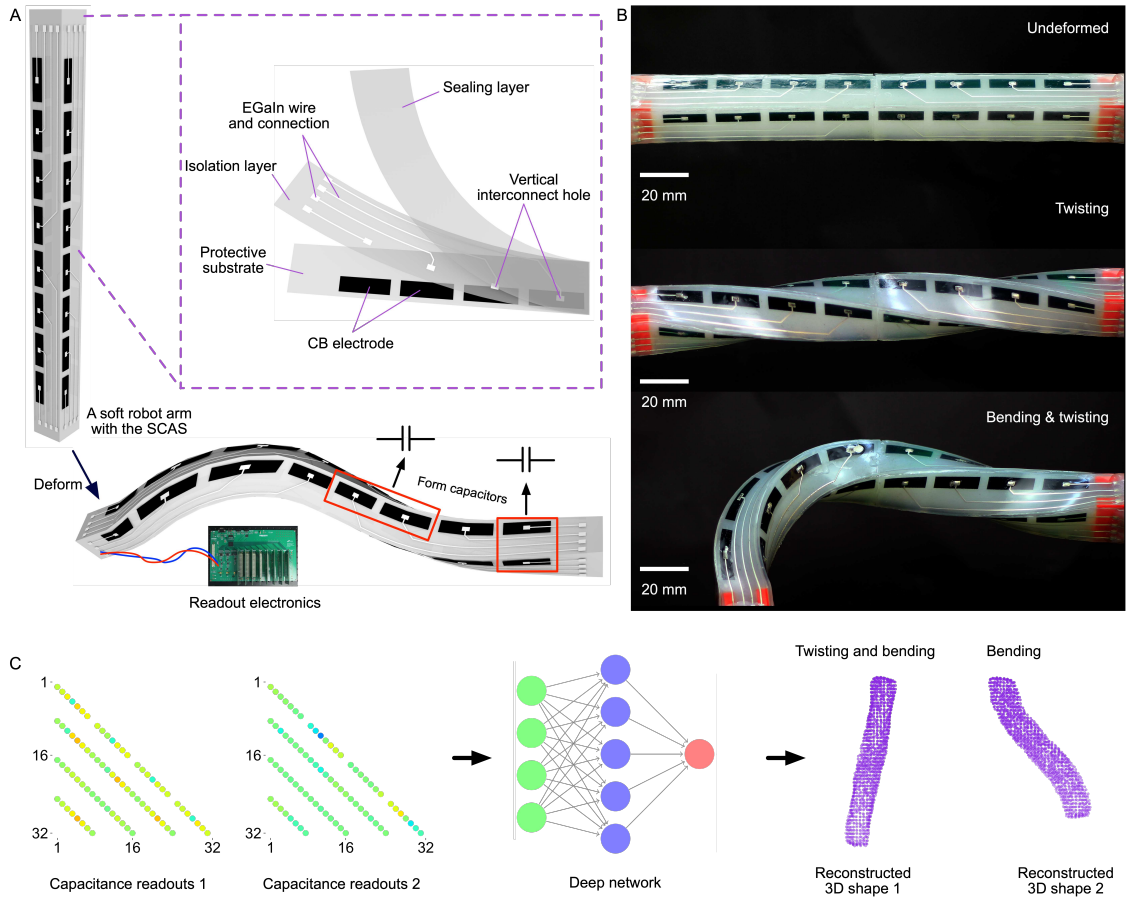


Figure 1: Design of the SCAS and the pipeline for full-geometry, high-resolution 3D deformation reconstruction of soft robots. (A) The entire SCAS that can cover the whole soft robot arm consists of multiple SCAS modules. Each module has 4 functional layers, i.e., the protective substrate (0.39 mm), the electrode layer (0.08 mm), the isolation layer (0.24 mm) and the sealing layer (0.3 mm). The soft electrodes are made of carbon black (CB) dispersed elastomers. Eutectic Gallium 75.5% Indium 24.5% (EGaIn) is employed to fabricate the wires and interfaces. The multi-position skin electrode combinations can form a sequence of capacitors. Geometric variations in the proximity of the electrode pair lead to the change of the corresponding capacitance. The readout electronics can record capacitance values of a selected set of electrode pairs at approximately 30 fps. (B) Snapshots of the soft arm in different states (undeformed, twisting and the combination of bending and twisting). (C) The data collected by the readout electronics is fed into a deep net and translated to a high-resolution representation of the full-body 3D shape of the soft arm.

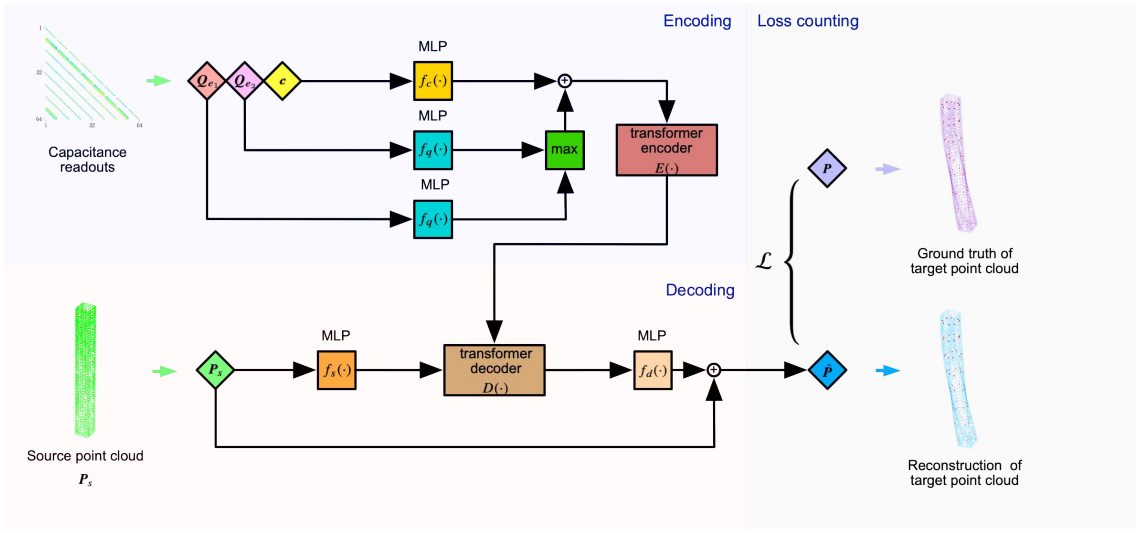


Figure 2: The architecture of the C2DT that infers the displacement of each point in the source point cloud (without any deformations) from the SCAS capacitance readouts. In the encoding part, the network encodes the input capacitance readouts and the geometrical structure information of electrode pairs to a high-dimensional space and feeds them to the transformer encoder to distil proprioceptive information. In the decoding part, the network manages to assign a correct displacement to each point in the source point cloud based on the output sequence of the encoding part. See Methods for more implementation and architecture details.

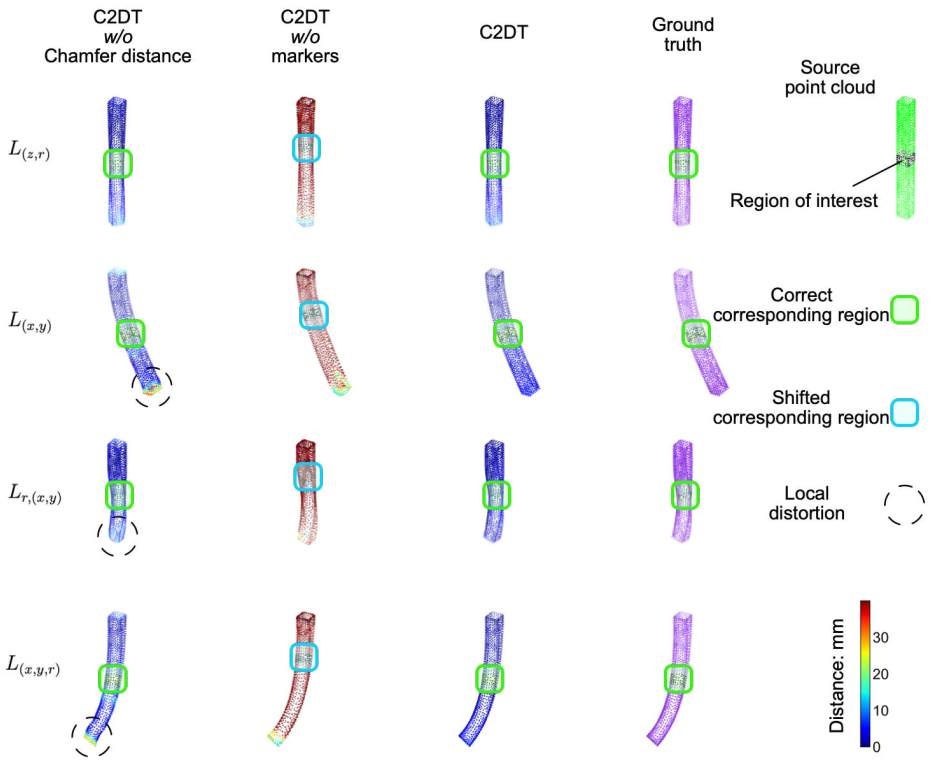


Figure 3: A set of examples of reconstructions generated by different C2DTs. The colour of each point in reconstructions indicates the distance from the corresponding point in the ground truth. The region of interest is the middle section in the source point cloud. The points in the region of interest (marked in black) should be mapped into the middle section in reconstructions if C2DTs learn correct point-to-point correspondences. We can observe apparent shifts in the reconstructions of the C2DT *w/o* markers.

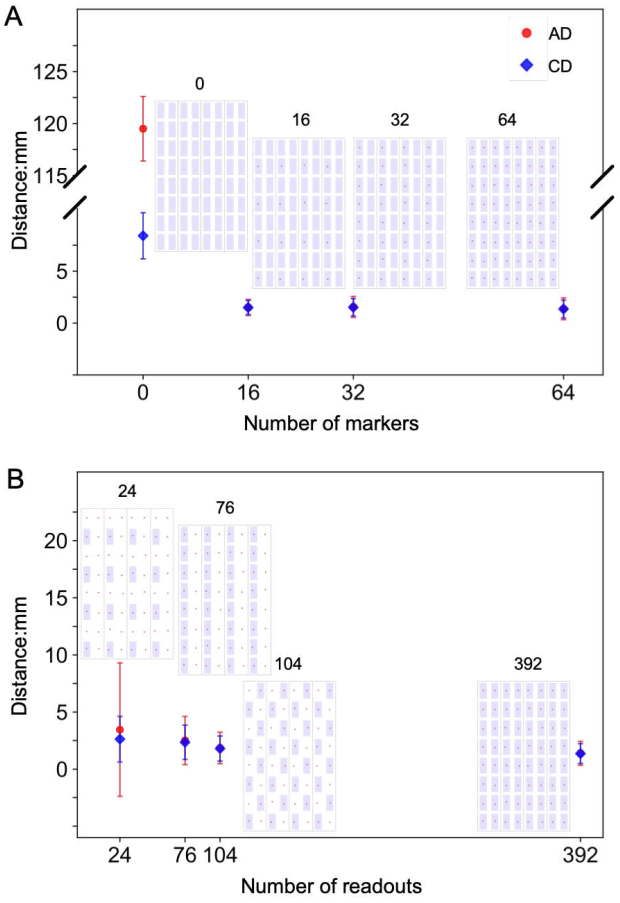


Figure 4: Reconstruction performance under different marker numbers and different electrode layouts. (A)The performance of the C2DT under 4 different numbers of markers (mean±standard deviation). (B) The performance of the C2DT under 4 different electrode layouts (mean±standard deviation).

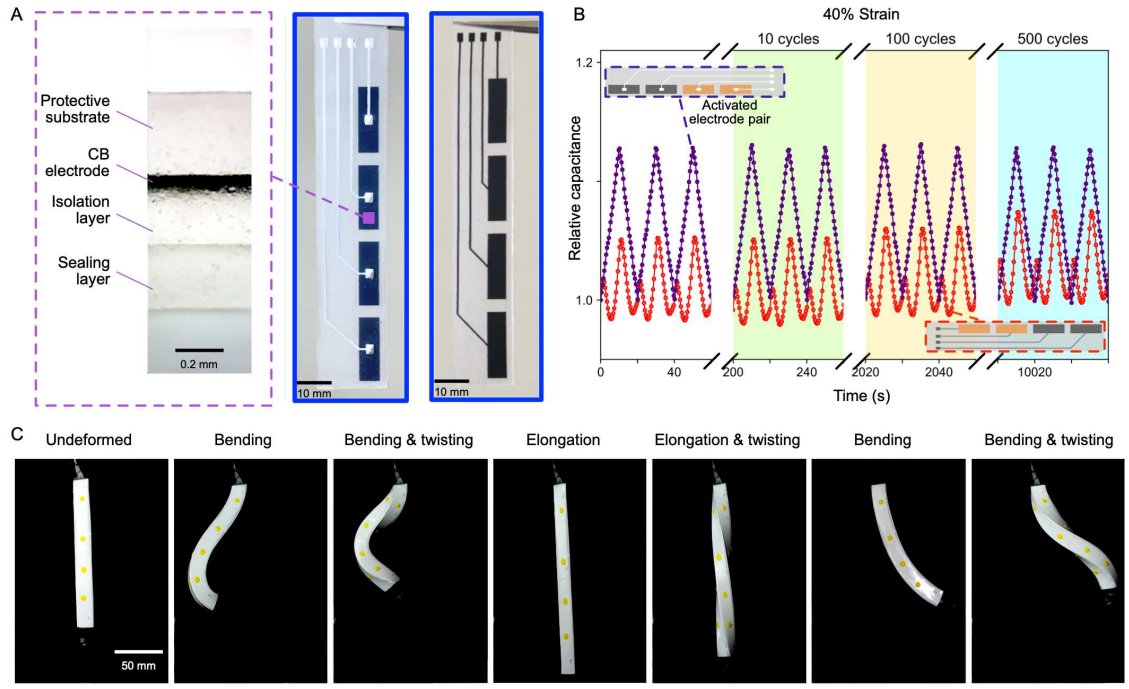


Figure 5: Characterization of the SCAS. (A) 4-electrode SCAS modules with EGaIn (centre) and CB (right) and wires. Left, the cross-section of the module with EGaIn wires under a 40x digital microscope. (B) Relative capacitance response curves of the two SCAS modules to a 40% periodic linear stretch. The SCAS with EGaIn wires shows better sensitivity, linearity and cycling stability than its CB wires counterpart. (C) A representative set of complex multimodal deformations that appear in our experiment.

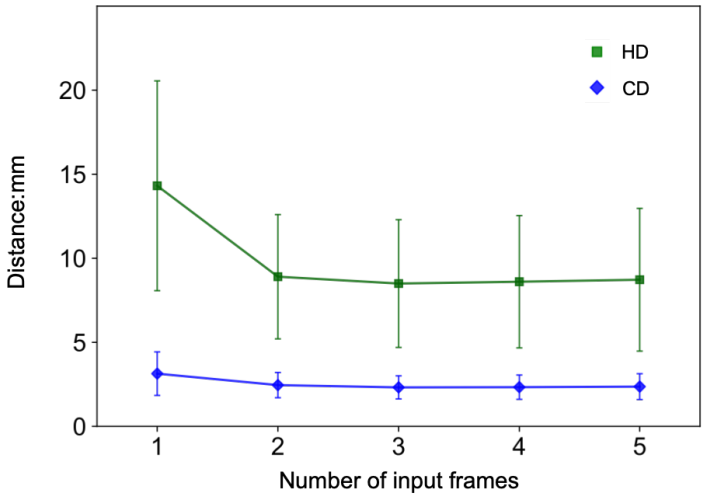


Figure 6: The performance of C2DTs, which take different numbers of adjacent frames as inputs, is shown (mean±standard deviation).

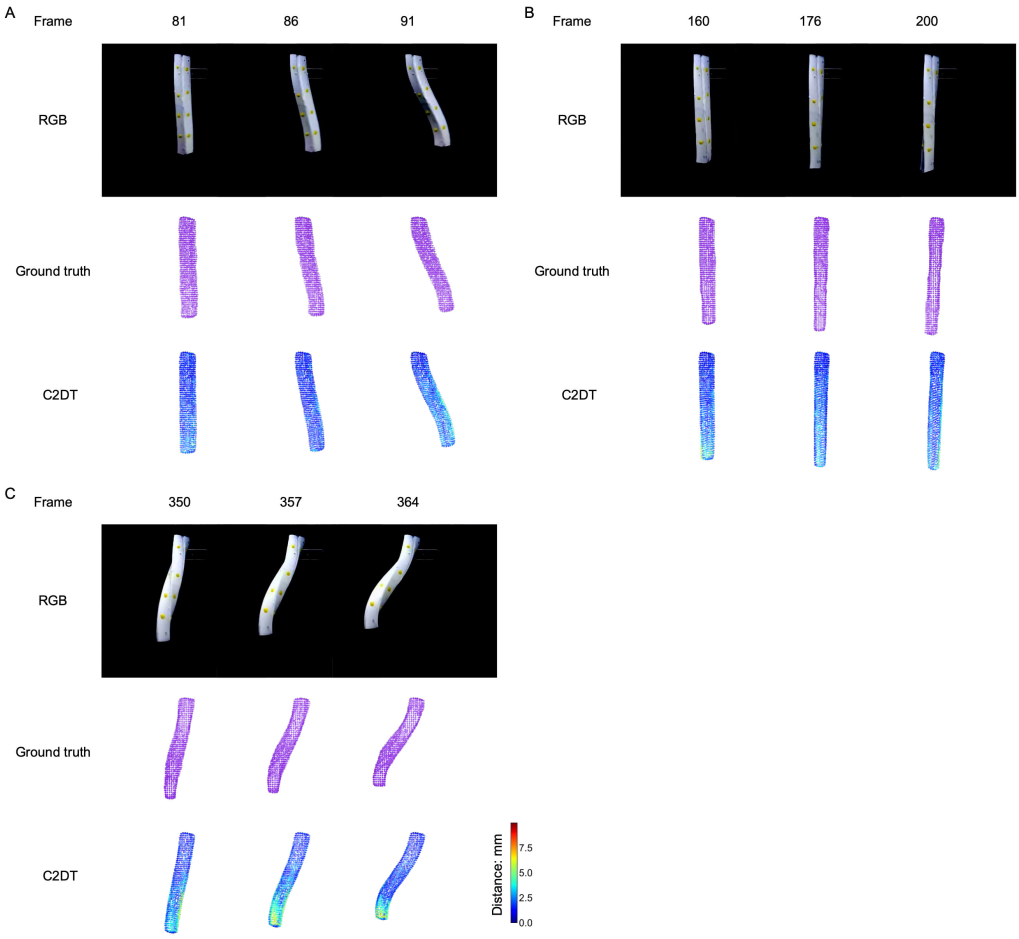


Figure 7: 3D full-geometry deformation reconstruction based on real-world SCAS signals. (A) Three representative frames of a bending process are shown in RGB images, ground truth point clouds collected by depth cameras and reconstructed point clouds of the C2DT. The colour of each point in reconstructions indicates the distance from the nearest point in the ground truth. Examples of other deformations are shown in (B) the combination of elongation and twisting and (C) the combination of bending and twisting.

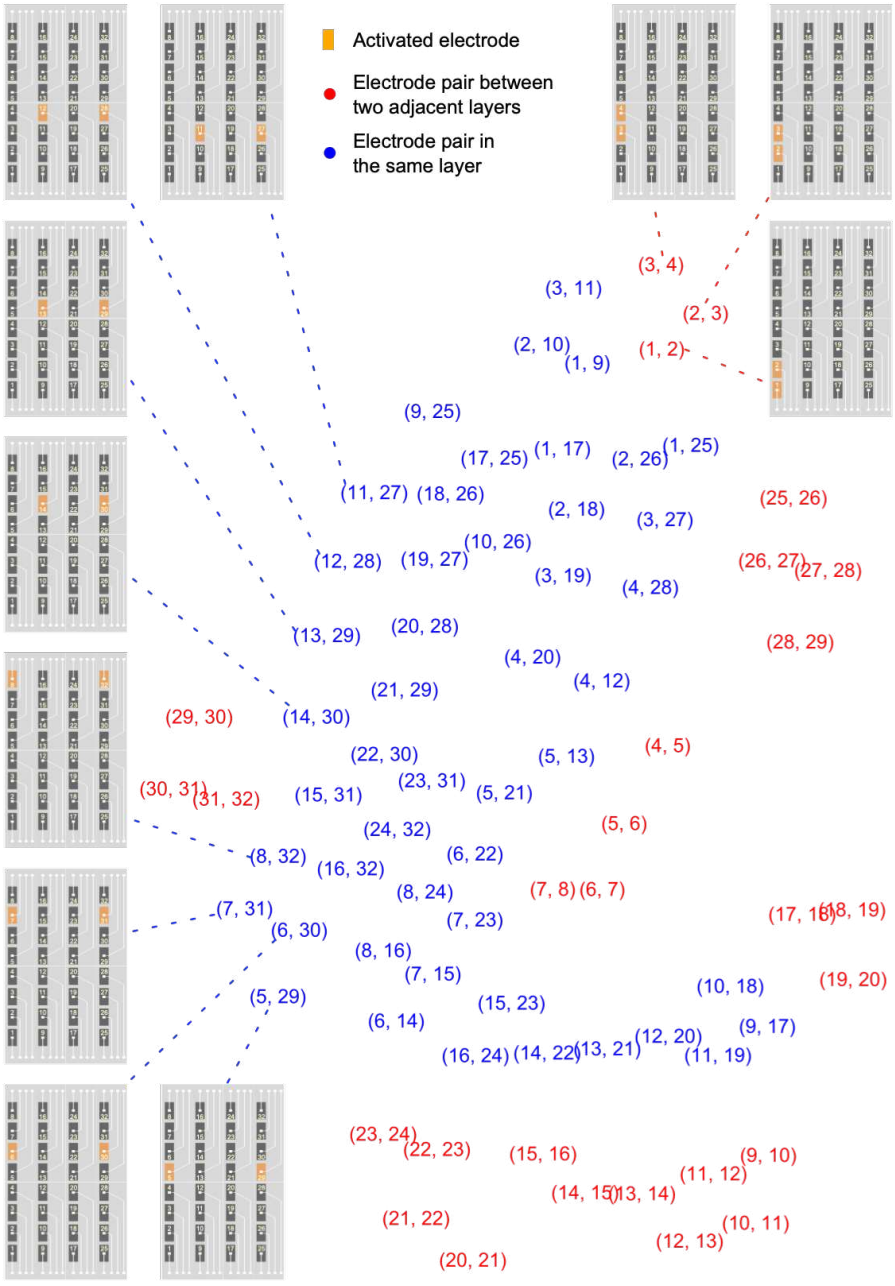


Figure 8: Two-dimensional t-SNE embedding of the output of the position encoding part in the trained C2DT. The electrode pairs between two adjacent layers and the electrode pairs in the same layer are marked in red and blue respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf