

Stretchable e-skin and transformer enable high-resolution morphological reconstruction for soft robots

Received: 30 June 2022

Delin Hu¹, Francesco Giorgio-Serchi^{2,3}, Shiming Zhang⁴ & Yunjie Yang  

Accepted: 30 January 2023

Published online: 23 February 2023

 Check for updates

Many robotic tasks require knowledge of the exact 3D robot geometry. However, this remains extremely challenging in soft robotics because of the infinite degrees of freedom of soft bodies deriving from their continuum characteristics. Previous studies have achieved only low proprioceptive geometry resolution (PGR), thus suffering from loss of geometric details (for example, local deformation and surface information) and limited applicability. Here we report an intelligent stretchable capacitive e-skin to endow soft robots with high PGR (3,900) bodily awareness. We demonstrate that the proposed e-skin can finely capture a wide range of complex 3D deformations across the entire soft body through multi-position capacitance measurements. The e-skin signals can be directly translated to high-density point clouds portraying the complete geometry via a deep architecture based on transformer. This high PGR proprioception system providing millimetre-scale, local and global geometry reconstruction (2.322 ± 0.687 mm error on a $20 \times 20 \times 200$ mm soft manipulator) can assist in solving fundamental problems in soft robotics, such as precise closed-loop control and digital twin modelling.

The neuro-proprioceptive system of animals mediates the perception of the body's geometry, constituting the prerequisite for precise and fast limb coordination during locomotion and interaction with the environment¹. Similarly, the dexterous manipulation of intelligent robots relies on the body's geometry estimation from the artificial proprioception system. Within the frame of conventional rigid robots, existing sensing technology already provides viable solutions to implementing body's geometry estimation that meets 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 degrees of freedom allow the full geometry to be defined by a bounded set of measurable parameters (such as joint angle and link length). However, the development of artificial proprioception systems for highly deformable structures, such as soft robots, remains a fundamental challenge,

severely restricting the understanding of soft robot behaviour and, ultimately, the capability to perform precise closed-loop control^{2,3}.

The highly deformable 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^{4,5}, thus highlighting their spontaneous vocation for biomedical applications⁶⁻⁸ and human–robot interaction⁹⁻¹¹ as well as their employment in unstructured, potentially cluttered scenarios. However, this very feature also gives infinite degrees of freedom to a soft body. It is infeasible to completely describe the three-dimensional (3D) morphology of a soft system with only a limited set of parameters^{4,12}. The number of independent parameters used by a soft proprioception system to describe the body geometry determines the smallest size of geometric

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

 e-mail: yyang@ed.ac.uk

variations that can theoretically be detected and presented by the system. Generally, the greater the number of independent parameters, the finer and more accurate the geometric variations can be described. Therefore, we define the number of such independent parameters as the proprioceptive geometry resolution (PGR). Soft proprioception systems with higher PGR are desirable for soft robotics as they can endow soft systems with bodily awareness more comparable to that of rigid robots, thus enabling more natural interaction with humans (for example, real-time 3D geometry of a soft robot can be visually observed without the line of sight, allowing users to operate robots intuitively even in occlusion environments) and underpinning precise closed-loop control.

To the best of our knowledge, there is no off-the-shelf high PGR soft proprioception system. Previous studies are focused on low PGR proprioception, limiting their capability to preserve geometric details (for example, local deformation and surface information) and their usage in practical application scenarios (for detailed comparison, see Supplementary Table 1)^{3,12–20}. For example, combined with the mathematical model of the soft robot under investigation, an optical fibre-based proprioception system can successfully reconstruct the 3D geometry on the basis of two parameters¹⁴, that is, global bending and twisting angles (PGR 2). However, the low PGR fails to describe local geometric variations (for example, bending of a robot segment) and is applicable only to a fixed bending direction and twisting axis (Supplementary Fig. 1). Some recent studies attempt to build soft proprioception systems with higher PGR by optimizing sensor design, introducing advanced machine learning algorithms (for example, long short-term memory (LSTM) networks) and employing 3D motion capture devices (for example, tracking cameras)^{3,12,17}. Redundant poly-dimethylsiloxane (PDMS) impregnated with conductive carbon nanotubes (cPDMS) sensors with LSTM can estimate 3D coordinates of a soft fingertip (PGR 3)¹². The simplified 3D geometry (described by nine parameters) of a trunk-shaped soft robot can be recovered through 12 conductive silicone-based piezoresistive sensors distributed on the robot body (PGR 9)³. The 3D deformation of a four-chamber pneumatic membrane (described by 49 visual markers) is reconstructed using LSTM and integrated optical sensors (PGR 147)¹⁷. Despite these recent advances, obtaining high PGR across a wide range of complex deformations remains unrealized.

In this Article, we propose a high PGR (3,900) proprioception system to confer full-geometry, millimetre-level bodily awareness to soft robots. The proprioception system encapsulates an intrinsically stretchable capacitive e-skin (SCAS) and a purpose-designed neural architecture (that is, the capacitance-to-deformation transformer, C2DT). Inspired by 3D electrical capacitance tomography (ECT)²¹, the SCAS has four different functional layers (Fig. 1a) and employs a redundant planar skin electrode layout (Fig. 1a,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 mechanism²² explores the dependency over the e-skin signals and directly translates the measurements to the point cloud of the morphology (Fig. 1c). The synergistic combination of the SCAS and C2DT can achieve accurate (2.322 ± 0.687 mm error on a $20 \times 20 \times 200$ mm soft manipulator) and high PGR (3,900; 1,300 points in each point cloud) 3D shape reconstruction under complex deformations, which is one or two orders of magnitude improvement over previous methods (for comparison, see Supplementary Table 1). The proposed system does not require mathematical modelling of the robot under investigation. Therefore, it theoretically should be agnostic to the shape of the soft body and has the potential to be extended to soft robotic platforms with an unprescribed morphology. This high PGR 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

Different from conventional parallel capacitive sensors frequently used in many previous studies^{2,18}, the design of SCAS is inspired by 3D ECT sensor and its sensing strategy²¹. 3D ECT has demonstrated that the capacitance readout of a boundary electrode pair is related to the permittivity of the medium within the sensitive region and its geometry. In soft robot proprioception, the permittivity remains constant. The change of capacitance primarily reflects geometric variations and, therefore, can be used to infer local and global deformations.

We first design the SCAS in the virtual environment and quantify its performance, before its physical implementation. 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 (Supplementary Fig. 2). 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. Supplementary Fig. 2a 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 (select two electrodes to form a capacitor, that is, $C_{64}^2 = 2,016$). We record only capacitances formed by electrodes in the same layer and those between two adjacent layers to ensure that they are practically measurable. Each SCAS measurement frame comprises 392 independent capacitance readouts (for measurement strategy details, see ‘Solid mechanics and electrostatics coupling field simulation’ in Methods and Supplementary Fig. 2). 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 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 (for details, see ‘Solid mechanics and electrostatics coupling field simulation’ in Methods). 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, that is, the compound deformation of elongation and twisting $L_{(z,\theta)}$, pure bending $L_{(x,y)}$, two-phase twisting and bending $L_{(x,\theta)}$ and the compound deformation of twisting and bending $L_{(x,y,\theta)}$. Simulation results show that SCAS signals can reflect the soft robot geometric variation under various complex deformations (Supplementary Fig. 3), indicating its feasibility as proprioceptors. We then leverage the virtual proprioception dataset to quantify the SCAS performance in high PGR full-geometry 3D deformation reconstruction. The results are utilized to optimize the design of the physical SCAS and learning-based proprioception algorithms.

C2DT

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 C2DT based on self-attention mechanism²² that is widespread in natural language processing^{23,24} and computer vision^{25,26} and shows superior performance in solving set-to-set problems. The framework of the C2DT is shown in Fig. 2a. C2DT infers the displacement of each point in the source point cloud (the one without deformation) from the proprioceptive

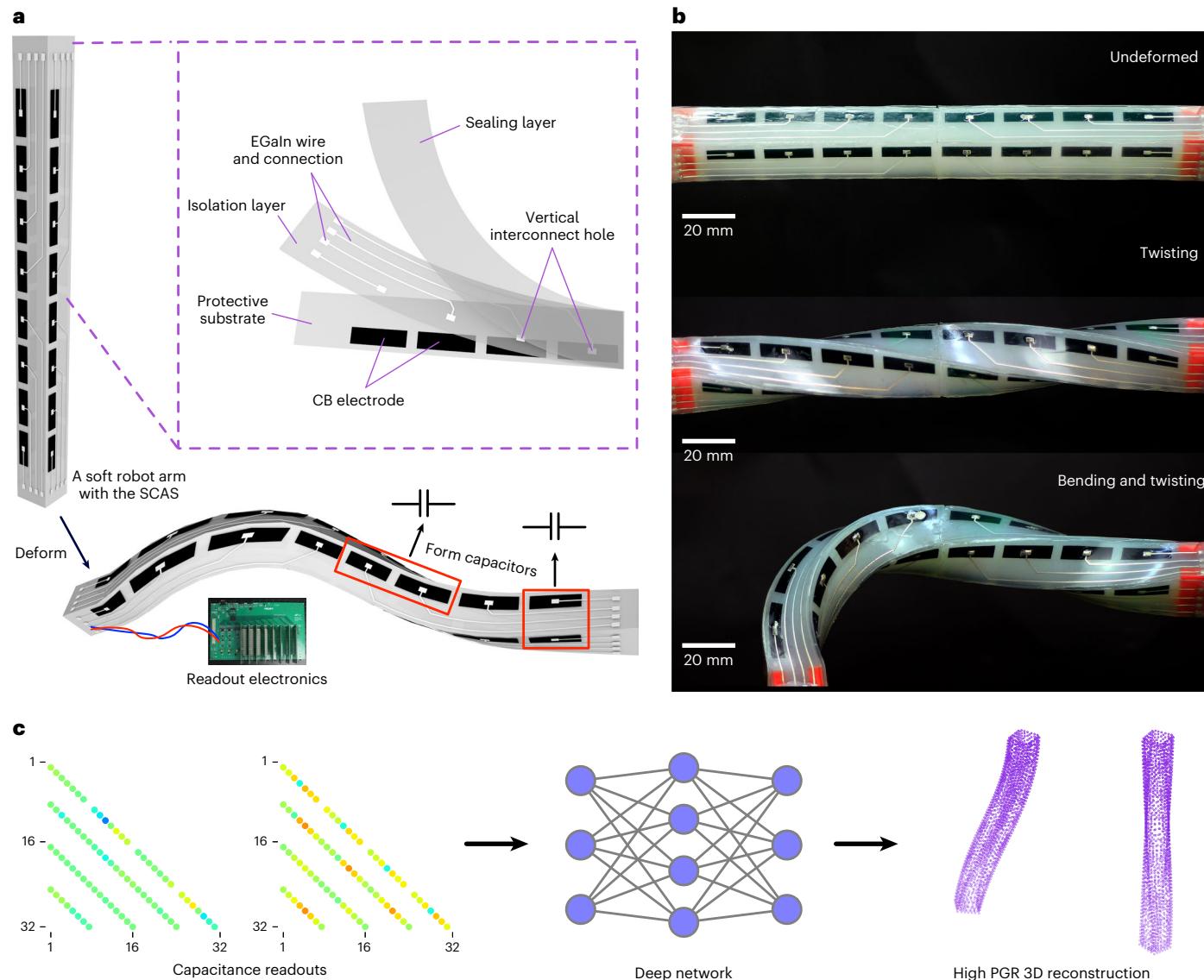


Fig. 1 | Design of the SCAS and the pipeline for full-geometry, high PGR 3D deformation reconstruction of soft robots. **a**, The entire SCAS that can cover the whole soft robot arm consist of multiple SCAS modules. Each module has four functional layers, that is, the protective substrate (0.39 mm), 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 CB dispersed elastomers. 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 in 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 compound deformation of bending and twisting). **c**, Data collected by the readout electronics are fed into a deep net and translated to a high PGR representation (point cloud) of the 3D robot shape.

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 readouts²⁷. We therefore design a special position encoding process in the C2DT to generate geometrical representations based on positions of individual electrode pairs (for more details, see Methods).

We train a C2DT using the virtual proprioception dataset by minimizing the loss function consisting of the squared distance term of visual markers (of which point-to-point correspondences are known) and the Chamfer distance (CD) term of the remaining points (of which point-to-point correspondences are unknown); for details of the loss function, visual markers and training, see Methods. The reconstruction results show superior PGR (that is, 5,148; 1,716 points represent the 3D geometry of the robot), accuracy (that is, 1.379 ± 1.048 mm;

Supplementary Table 2) and are able to capture the whole range of complex deformations tested (Fig. 2b, Supplementary Fig. 3 and Supplementary Video 1). We employ four error metrics to quantitatively evaluate the performance of C2DT, that is, the average distance (AD), the maximal distance (MD), the CD²⁸ and the Hausdorff distance²⁹ (HD); for expressions of these metrics, see Methods. We train several C2DTs with different hyperparameters and compare their performance to determine an optimized network structure. The quantitative results are presented in Supplementary Table 2. We find that the C2DT with six transformer layers outperforms the other candidates. The AD error achieved with this setup is as low as 1.379 ± 1.048 mm, comparable to the accuracy achieved with RGB-D cameras frequently used as ground truth in the relevant research²⁹.

We implement ablation studies of the C2DT with six transformer layers to better understand the role of each loss term and position encoding. The results are shown in Fig. 2b, Supplementary Fig. 3 and

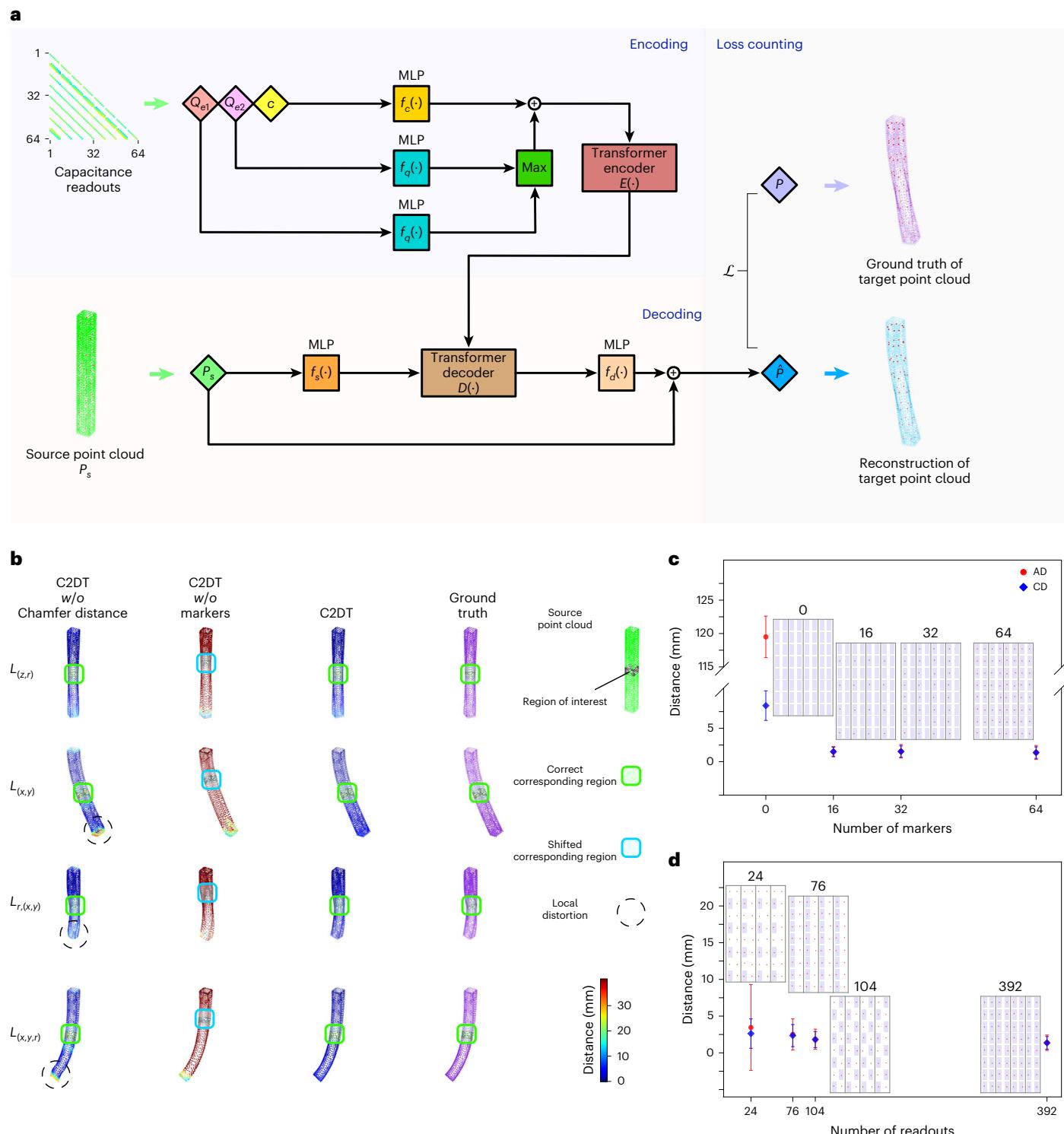


Fig. 2 | High PGR 3D deformation reconstruction based on the virtual dataset.

a, The architecture of the C2DT that infers the displacement of each point in the source point cloud (without deformation) 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. For more implementation and architecture details, see Methods. **b**, 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. **c**, The performance of the C2DT under four different numbers of markers (mean \pm standard deviation on 7,096 testing samples). **d**, The performance of the C2DT under four different electrode layouts (mean \pm standard deviation on 7,096 testing samples).

Supplementary Table 3. We observe that the C2DT cannot learn correct point-to-point correspondences without including visual markers in training. This phenomenon is illustrated in Fig. 2b, where the points in the region of interest of the source point cloud are not mapped into the correct corresponding region using the C2DT without (*w/o*) markers. Although reconstructions show similarities with the ground truth by minimizing the CD 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 CD 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*-distributed stochastic neighbour embedding (*t*-SNE)³⁰ (Supplementary Fig. 4) and see that, after position encoding, the electrode pairs with high geometrical correlation tend to cluster together, and the electrode pairs 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 pairs.

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 of the physical system. 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. 2c, 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 (for example, 32), as illustrated in Fig. 2d. These results highlight a favourable trade-off between reconstruction accuracy and electrode/markers units, confirming that it is safe to sacrifice a minute part of performance to simplify the fabrication and deployment of the SCAS.

Fabrication and characterization of the e-skin

On the basis of the conclusions from the above investigation, we design a physical SCAS with 32 electrodes, consisting of eight 4-electrode SCAS modules. This design balances the full-geometry reconstruction performance with fabrication complexity. We fabricate multiple four-electrode SCAS modules in parallel using established elastomer processing technologies³¹. The electrodes are made of carbon black (CB) dispersed elastomers. However, this material is unsuitable for wires and interfaces due to its high resistivity and non-linear, irreversible conductivity response under deformation^{32,33}. Therefore, eutectic gallium (75.5%)-indium (24.5%) (EGaIn) is employed to fabricate the wires and interfaces due to its high conductivity ($3.4 \times 10^7 \text{ S m}^{-1}$) and stable response to deformation. The fabrication process is presented step by step in Supplementary Fig. 5a, and additional details regarding materials and fabrication are reported in Methods.

The four-electrode SCAS module (20 × 120 mm) consists of four different functional layers (Figs. 1a and 3a), that is, the protective substrate (thickness 0.39 mm), the electrode layer (0.08 mm), the isolation layer (0.24 mm) and the sealing layer (0.3 mm). We engrave microchannels for wires (width 0.5 mm) 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 microchannels. 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 Supplementary Fig. 7a). The response curves show excellent linearity and consistency over multiple cycles (more than 500 cycles

in Supplementary Fig. 7b,c). For comparison, we characterize a SCAS with CB wires using the same approach (Supplementary Fig. 7d,e). As Fig. 3b 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 eight four-electrode SCAS modules on the surface of a soft manipulator with the size of 20 × 20 × 240 mm (Supplementary Fig. 5b; 40 mm in height for the interface area, that is, not reconstructed). The 32-electrode SCAS, consisting of eight SCAS modules, connects to an in-house developed data acquisition system³⁴ 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 (Supplementary Fig. 5c). The experiment platform (Supplementary Fig. 8) can synchronously record capacitance and point cloud data at a frame rate of around 30 frames per second (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 seconds (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 omnidirectional bending, omnidirectional elongation, twisting around an arbitrary axis and their compound deformations (Fig. 3c), during the experiment. In most frames, point clouds collected by cameras cannot represent full-geometry 3D deformations due to missing points caused by inevitable visual occlusion. We fill points by α shape reconstruction³⁵ for the frames with minor missing point issues and directly filter the frames that have severe occlusion. Then we obtain a total of 30,973 frames of data (for details of data acquisition and pre-processing, see Methods). A set of samples in this dataset is shown in Fig. 4a and Supplementary Fig. 9.

Real-world high PGR proprioception

The challenge of real-world high PGR 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. To compensate for these added sources of inaccuracy, we enhance the C2DT framework by increasing the number of input frames (N , 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 three adjacent input frames (Fig. 4b). This improvement indicates that increasing the number of input frames can reduce the negative impacts of noise in SCAS signals and synchronization between devices. The temporal correlation among adjacent frames can also be considered favourable for deformation reconstruction. A representative set of reconstructions of the C2DT with three input frames is shown in Fig. 4c and Supplementary Video 2. The results achieve $2.322 \pm 0.687 \text{ mm}$ for the CD metric (Supplementary Table 4) with the PGR of 3,900 (that is, 1,300 points in each point cloud). Compared with simulation results, some elaborate geometrical features of reconstructed deformations in certain frames are less obvious, especially for those related to twisting (Supplementary Video 3). This is mainly because the ground truth point clouds acquired

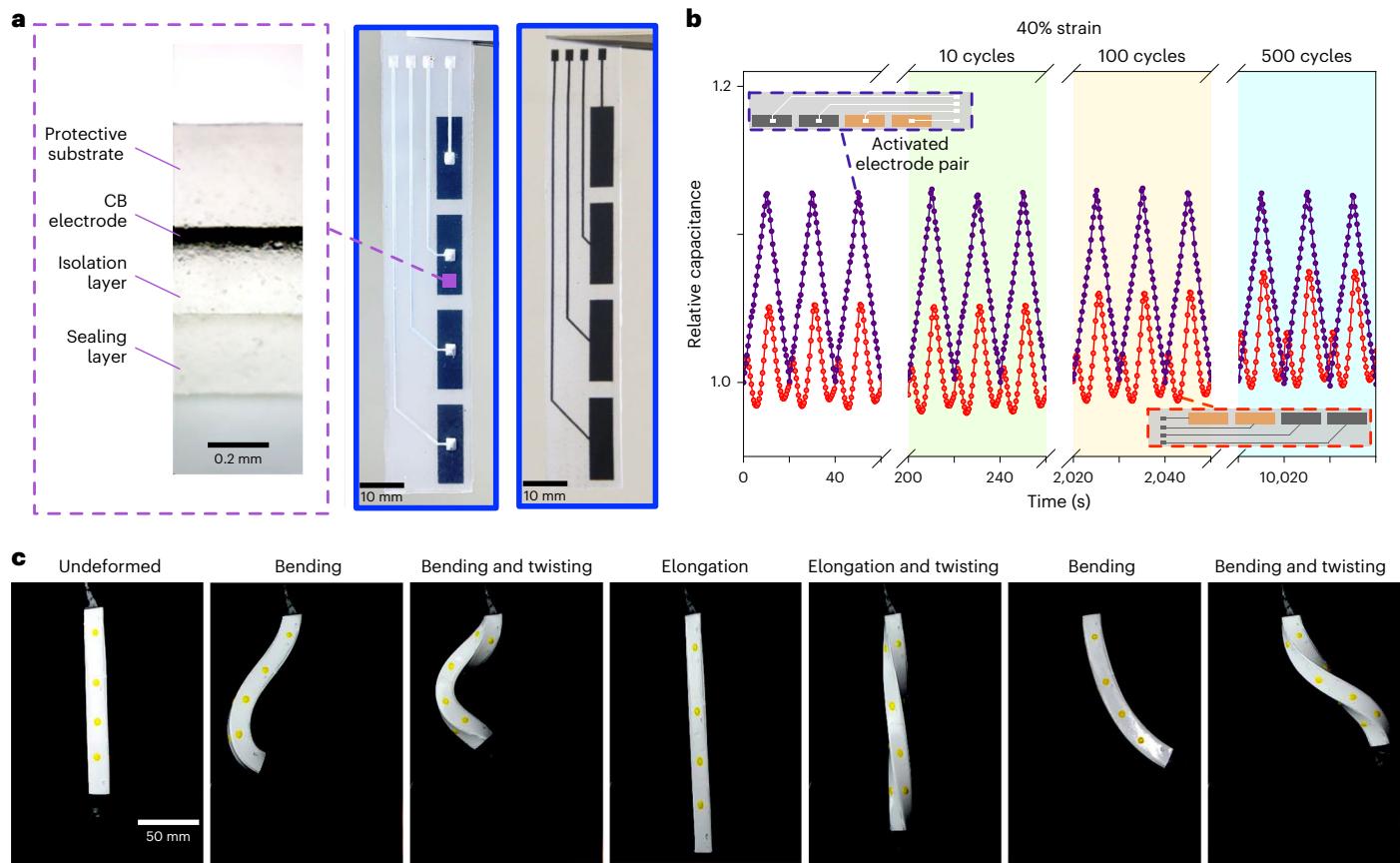


Fig. 3 | Characterization of the SCAS. **a**, Four-electrode SCAS modules with EGaIn (centre) and CB (right) wires. Left: the cross-section of the module with EGaIn wires under a $40\times$ digital microscope. **b**, Relative capacitance response curves of the two SCAS modules to a 40% periodic linear stretch. The SCAS with

EGaIn wires shows better sensitivity, linearity and cycling stability than its CB wires counterpart. **c**, A representative set of complex deformations that appear in our experiment.

by RGB-D cameras cannot reach the quality of point clouds synthesized by 3D-SECF simulation.

According to the ablation study (Supplementary Fig. 10), 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 reconstruction quality (Supplementary Table 4). The position encoding part is crucial to extract useful proprioceptive information from physical SCAS signals. Similar to its contribution in the training with the simulation dataset, position encoding can assign discriminative high-dimensional representations to different electrode pairs based on their geometrical structures (Supplementary Fig. 11).

Discussion

We presented a proprioception system that could visualize high PGR 3D full-geometry deformations of soft robots. It is empowered by an intrinsically stretchable 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 proprioception system on complex deformations, several issues remain to be addressed to fully exploit its potential.

The SCAS fabrication involves manual operation (for example, 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. Furthermore, the thickness of the SCAS is about 1 mm, which is suitable for demonstrating high PGR proprioception in our soft robot testbed ($20 \times 20 \times 200$ mm) and other similar proprioception scenarios. However, more advanced fabrication approaches^{36–38} could be adopted to extend the proposed framework to other application domains, such as skin-interfaced wearable devices.

While this work focuses on proprioception induced by external forces applied to the tip of the robot, real-world operation entails many other kinds of stimuli from the environment. These stimuli, in turn, may be the source of soft body deformations (for example, compression) and involve peripheral information (for example, temperature, texture). Due to the capacitive nature of SCAS, it can, in principle, simultaneously detect several distinct types of external stimuli, such as tactile mapping and permittivity of the objects in the proximity of the robot. The SCAS signals could be interfered, and the accuracy of the shape reconstruction might drop when external stimuli occur. Two possible solutions for this issue are envisaged. First, a stretchable conductive layer that is grounded can be integrated to the top of the SCAS, shielding the external electrical interference. In addition, sensitivity to external stimuli provides an opportunity to measure them. More advanced data interpretation algorithms could be developed to decouple deformation and external stimuli information from SCAS signals, which however could be highly challenging. Integrating multi-modal sensors into the SCAS framework has the potential to alleviate this issue and enhance the capability to detect multiple external stimuli simultaneously.

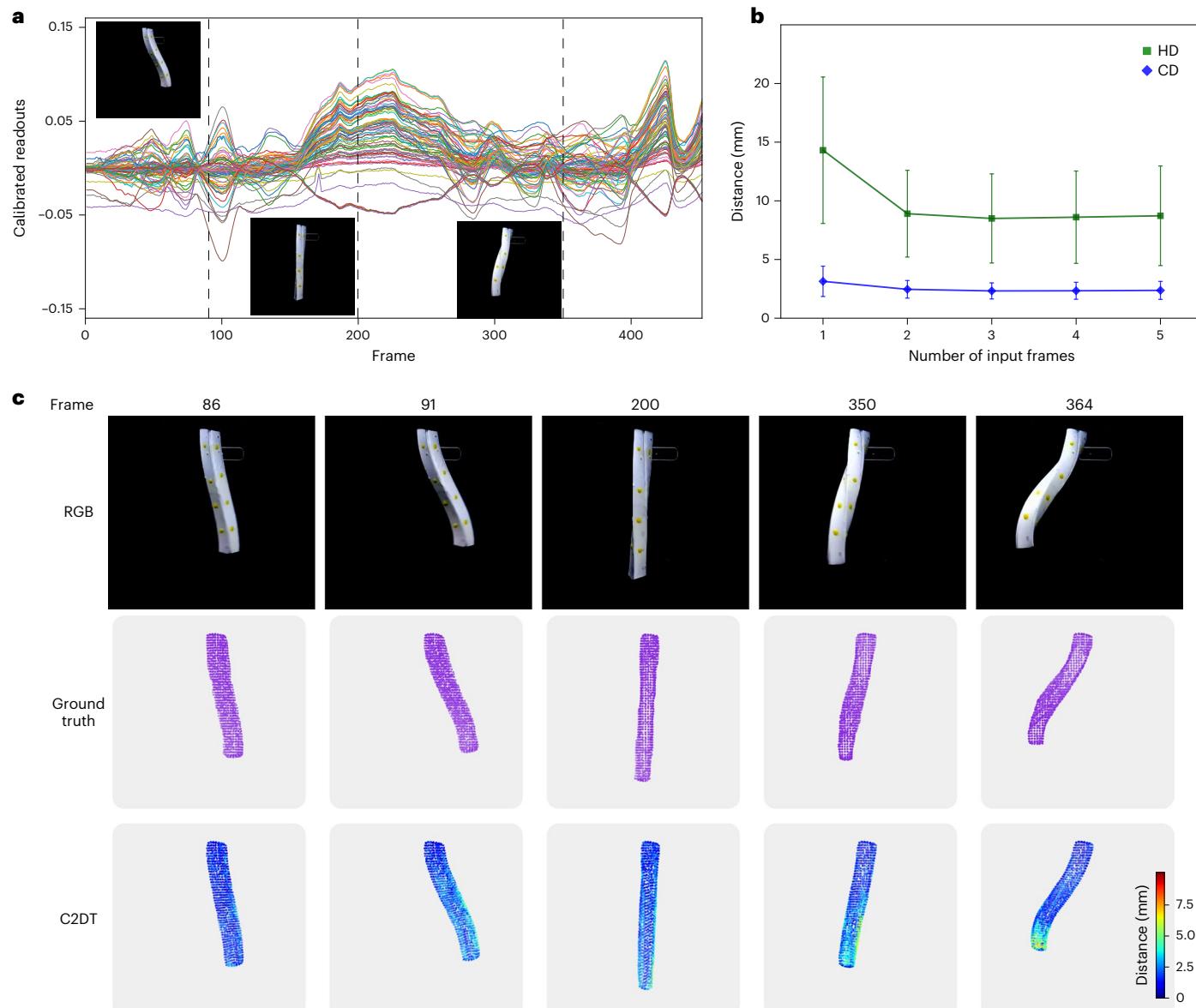


Fig. 4 | Real-world high PGR proprioception. **a**, The curves of calibrated capacitance readouts of the SCAS during a period of about 20 s in the experiment. Each readout is calibrated as follows: $c = (c' - c_{\text{emp}})/c_{\text{emp}}$, where c is the calibrated capacitance readout, c' is the original readout and c_{emp} is the

readout without deformation. **b**, The performance of C2DTs, which take different numbers of adjacent frames as inputs (mean \pm standard deviation on 4,262 testing samples). **c**, A representative set of examples of high PGR 3D deformation reconstruction.

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 deformations 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 labelled training samples. However, the gap between virtual and physical environments leads to performance deterioration if the network is trained only on the simulation dataset. Sim-to-real transfer learning methods are considered as the potential solution. The development and application of sim-to-real approaches suitable for soft robot proprioception can substantially increase the value of the simulation data and reduce the cost of real-world data acquisition.

Notwithstanding the above limitations, the proposed proprioception system can achieve real-time (30 fps), high PGR (3,900)

full-geometry deformation reconstruction with high accuracy (2.322 ± 0.687 mm CD error) under complex deformations. This level of proprioception represents a step change over previous attempts and is beyond existing proprioception systems. Notably, the system has the potential to be extended to different types of soft bodies through a straightforward learning process without requiring a priori knowledge. Implementing such high PGR, 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.

Methods

Solid mechanics and electrostatics 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 1,000 mm; Supplementary Fig. 2a). 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 $\nu = 0.022$, density $\rho = 1.28 \times 10^3$ kg m $^{-3}$, relative permittivity $\epsilon_r = 3$.

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 compound deformation 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 compound deformation 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 Supplementary Fig. 3. Since it is impractical to ascertain the exact point-to-point correspondences of all points between two different deformations in the physical world, we resort to a scheme that can be realistically implemented. We select 64 points as visual markers, whose correspondences are available during network training and the correspondences of the remaining points are only used in testing for evaluation (Supplementary Fig. 2a).

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. Supplementary Fig. 2b shows all 28 electrode pairs in the first layer that form measurable independent capacitors. Supplementary Fig. 2c 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 Supplementary Fig. 3.

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. 2a) that is able to deform the source point cloud \mathbf{P}_s to approximate the target point cloud \mathbf{P} based on the measurement characteristic tensor ($\mathbf{c}, \mathbf{Q}_{e_1}, \mathbf{Q}_{e_2}$). Here, $\mathbf{P}_s \in \mathbb{R}^{N_p \times 3}$ is the point cloud without deformation; N_p is the number of points in \mathbf{P}_s , which is 1,716 in this case; $\mathbf{P} \in \mathbb{R}^{N_p \times 3}$ and $\hat{\mathbf{P}} \in \mathbb{R}^{N_p \times 3}$ are the ground truth and reconstructed point clouds with a specific deformation, respectively; $\mathbf{c} \in \mathbb{R}^{N_m \times 1}$ is the corresponding calibrated capacitance readouts vector; N_m is the number of readouts in \mathbf{c} with the value of 392 in this case; and $\mathbf{Q}_{e_1} \in \mathbb{R}^{N_m \times 3}$ and $\mathbf{Q}_{e_2} \in \mathbb{R}^{N_m \times 3}$ are the coordinates of electrodes to generate \mathbf{c} .

The C2DT architecture consists of two parts, that is, encoding and decoding. The input of the encoding part is $\mathbf{c}, \mathbf{Q}_{e_1}$ and \mathbf{Q}_{e_2} . \mathbf{Q}_{e_1} and \mathbf{Q}_{e_2} are considered as positional signals that can help distinguish different elements in \mathbf{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 \mathbf{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, \mathbf{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 $\hat{\mathbf{P}}$ is obtained by adding it to \mathbf{P}_s .

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

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

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

The structures of subnetworks of the C2DT are as follows:

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

where $h_{\text{em}} = 32$, $d_{\text{model}} = 128$, $a = 1.2$, $d_{\text{ff}} = 256$, $h = 8$, $P_{\text{drop}} = 0.1$, $n_{\text{e-layer}} = 3$, $n_{\text{m-layer}} = 1$ 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 transformer²². We remove the first self-attention cell of Transformer.DecoderLayer and use the remaining part as Transformer.MutualLayer because \mathbf{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, that is, 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 testing 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 $L_{(x,y,r)}$.

Quantifying the range of deformations can assist in evaluating the reconstruction performance. However, it is challenging to characterize the range of complex deformations using only several parameters,

such as bending angle and/or elongation displacement. Otherwise, a low PGR proprioception system would be sufficient to provide accurate geometry reconstruction. Here we characterize the deformation range using (1) the range of coordinates of points and (2) the maximum displacement of the centroid. In the simulation, the coordinates of testing samples are in the range of -724.27 to 728.68 mm in the *x* direction, -742.15 to 743.66 mm in the *y* direction and -728.09 to 487.95 mm in the *z* direction. Note that we set the centroid of the point cloud without deformation as the origin. The maximum displacement for the centroid is 341.92 mm.

The C2DT is implemented in Python and PyTorch packages³⁹. We use the Adam⁴⁰ optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$) 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 $\lambda = \max(1, 300 - 2 \times (\text{epoch} - 1))$. 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 four error metrics, that is, the AD, the MD, the CD and the HD:

$$\text{AD} = \frac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{p}^i - \hat{\mathbf{p}}^i|_2 \quad (2)$$

$$\text{MD} = \max_{i \in [1, N_p]} |\mathbf{p}^i - \hat{\mathbf{p}}^i|_2 \quad (3)$$

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

$$\text{HD} = \max \left(\max_{\mathbf{p} \in \mathcal{P}} \min_{\mathbf{p} \in \mathcal{P}} |\mathbf{p} - \hat{\mathbf{p}}|_2, \max_{\mathbf{p} \in \mathcal{P}} \min_{\mathbf{p} \in \mathcal{P}} |\mathbf{p}^i - \hat{\mathbf{p}}|_2 \right) \quad (5)$$

We compare the performance of C2DTs with different hyperparameters, and the results are presented in Supplementary Table 2. 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 CD term, respectively, and perform the same training procedure to obtain results of the C2DT *w/o* markers and the C2DT *w/o* CD. The reconstructed point clouds and values of these metrics are shown in Fig. 2b, Supplementary Fig. 3 and Supplementary Table 3. 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 are visualized through t-SNE³⁰ and presented in Supplementary Fig. 4, 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. 2c,d.

SCAS fabrication, characterization and deployment

The 32-electrode SCAS comprises eight modular 4-electrode SCASs. Each SCAS module has four different functional layers, that is, 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 Supplementary Fig. 5a: (1) 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 micrometre film applicator to flatten the silicone and cure it for 3 min at 100 °C. (2) We first mix Imerys Enasco 250P conductive CB (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. (3) 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 × 6 mm, which is one-fifth of the one studied in the simulation. (4) We use the same method as in step 1 to fabricate a silicone membrane for the isolation layer on the top of the electrode layer. (5) Two rounds of engraving are performed with 20.5% power, 300 mm s⁻¹ speed and 0.05 mm interval to generate microchannels 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 × 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. (6) We fabricate a new silicone membrane following step 1, 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 5 with the membrane. The curing takes about 4 h under room temperature to ensure high-quality bonding. (7) We inject EGaIn (Sigma Aldrich) ink from readout connections, and meanwhile exhaust the air in microchannels through the vertical interconnect holes. (8) We obtain the final modular four-electrode SCAS. The planar size of the SCAS module is 120 × 20 mm, of which 100 × 20 mm is the area of the electrodes, and 20 × 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 five 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 four-electrode SCAS with CB wires and a four-electrode SCAS with EGaIn wires on the front and back sides of a segment of the square cylinder silicone structure (20 × 20 × 140 mm) and cyclically stretch them using a Nema23 stepper motor with a SFU1605 ball screw (Supplementary Fig. 7a). 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 Supplementary Fig. 7b–e. 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 (Ecoflex 00-30) with the size of 20 × 20 × 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 eight 4-electrode SCAS modules on its surface to form the 32-electrode SCAS (Supplementary Fig. 5b). The soft robot and SCAS are fabricated with the same material (Ecoflex 00-30), which allows them to be firmly merged, with no modulus mismatch, by using uncured Ecoflex 00-30 silicone as adhesive. The unity of the material enables the robot and SCAS to be considered as a whole system during experiments, thus minimizing the effect of SCAS on the original robot motion and deformation. Supplementary Fig. 6 shows the adhesion between the SCAS and the robot under various deformations. No separation or dislocation was observed in all cases. The transparency of silicone adversely impacts 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 (Supplementary Fig. 5c). Individual electrodes are indexed and accessible from the readout electronics.

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 cameras⁴¹ and a laptop installed with a customized software to control the readout electronics and record data from the cameras and the SCAS (Supplementary Fig. 8). The readout electronics is based on a 32-electrode ECT system that supports arbitrary switching schemes³⁴. 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 pre-processing

In real-world experiments, we manually manipulate the hand holder bonding to the bottom of the robot arm to induce a variety of complex deformations, including omnidirectional bending, twisting around an arbitrary axis, omnidirectional elongation and their compound deformations (Fig. 3c and Supplementary Fig. 9b,c). Meanwhile, we synchronously record the SCAS and camera data (that is, 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 20 s) 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^{42,43}. The raw data are noisy and contain many meaningless background points, making them unusable for direct training. We clean and pre-process the data using MATLAB 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. To further reduce the negative impact of noise and outliers, we filter out regions whose local point densities are lower than a pre-set threshold. Due to inevitable visual occlusion, in many frames the cleaned point clouds cannot completely represent 3D deformations. To alleviate this issue, further pre-processing is required before training. We implement average grid downsampling with a 4 mm box gird filter at first for computational efficiency. Then we reconstruct α shapes³⁵ 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. To meet this requirement, we first implement average grid downsampling with a 4 mm box gird filter and then use farthest point sampling⁴⁴ 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 ten 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 four 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

axis of the fused coordinate system among all four 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 two markers in all layers are retained while others are dismissed.

Upon the above filtering process, a total of 30,973 frames of data remain 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.

$$\begin{aligned} \mathcal{L}^* = \mathbb{E}_{P \sim \mathcal{P}} \left\{ \lambda_1 \sum_{k=1}^{N_t} \sum_{l=1}^{N_w} \left[d(\hat{\mathbf{p}}_{l_k}^i, \mathbf{P}_{l_k}) \cdot S_{r2g}^{k,i} + d(\mathbf{p}_{l_k}^i, \hat{\mathbf{P}}_{l_k}) \cdot S_{g2r}^{k,i} \right] \right. \\ \left. + \lambda_2 \sum_{j=1}^{N_r} \left[d(\hat{\mathbf{p}}_r^j, \mathbf{P}_r) + d(\mathbf{p}_r^j, \hat{\mathbf{P}}_r) \right] \right. \\ \left. + \lambda_3 \sum_{j=1}^{N_r} \sum_{l=1}^{N_w} \left[\left(\|\mathbf{p}_r^j - \hat{\mathbf{p}}_r^{j,l} \|_2 - \delta_d \cdot s^{j,l} \right)^2 \cdot S_d^{j,l} \right. \right. \\ \left. \left. + \left(\|\mathbf{p}_r^j - \hat{\mathbf{p}}_r^{j,l} \|_2 - \delta_u \cdot s^{j,l} \right)^2 \cdot S_u^{j,l} \right] \right\} \end{aligned} \quad (6)$$

The first term of \mathcal{L}^* counts the CD between the reconstruction and the ground truth of markers layer by layer, where $\mathbf{P}_{l_k} \in \mathbb{R}^{N_w \times 3}$ is the coordinates of the visual markers in the l_k layer; $\mathbf{p}_{l_k}^i \in \mathbb{R}^3$ is the coordinates of the i th point in \mathbf{P}_{l_k} ; $d(\hat{\mathbf{p}}_{l_k}^i, \mathbf{P}_{l_k})$ is the squared distance between $\hat{\mathbf{p}}_{l_k}^i$ and the nearest point in \mathbf{P}_{l_k} ; N_t is the number of layers; N_w is the number of markers in each layer and the values of N_t and N_w are 4 in this case. When computing the loss, we only need to consider the marker points extracted in the data pre-processing and ignore the padding points. Note that all points in $\hat{\mathbf{P}}_{l_k}$ 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. To eliminate the effect of padding points during training, we synthesize masks $S_{r2g}^{k,i}$ and $S_{g2r}^{k,i}$ as follows.

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

The second term in \mathcal{L}^* is exactly the same as its simulation counterpart that counts the CD 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 dramatically before and after deformations, where $\hat{\mathbf{p}}_r^{j,l}$ is the l th neighbour of \mathbf{p}_r^j , $s^{j,l}$ 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 pre-set range. We achieve this with masks $S_d^{j,l}$ and $S_u^{j,l}$ as follows.

- $S_d^{j,l}$ is set to 1 if $\|\mathbf{p}_r^j - \hat{\mathbf{p}}_r^{j,l}\|_2 - \delta_d \cdot s^{j,l} < 0$, otherwise $S_d^{j,l}$ is set to 0.
- $S_u^{j,l}$ is set to 1 if $\|\mathbf{p}_r^j - \hat{\mathbf{p}}_r^{j,l}\|_2 - \delta_u \cdot s^{j,l} > 0$, 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_t) adjacent frames as its input.

The first linear cell in f_c is therefore modified to $\text{Linear}(N_i, h_{\text{em}})$. The hyperparameters of the C2DT are set as: $h_{\text{em}} = 32$, $d_{\text{model}} = 64$, $d_{\text{ff}} = 128$, $h = 4$, $P_{\text{drop}} = 0.1$, $n_{\text{e-layer}} = 2$, $n_{\text{m-layer}} = 1$ and $n_{\text{d-layer}} = 1$. 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 26,711 frames (about 1,020 s) are used for training (20,693 frames) and validation (6,018 frames), and the last 4,262 frames (about 200 s) are used for testing. The coordinates of testing samples are in the range of -141.01 to 129.99 mm in the x direction, -98.41 to 190.91 mm in the y direction, and -100.28 to 111.73 mm in the z direction (the centroid of the point cloud without deformation is set as the origin). The maximum displacement for the centroid is 72.38 mm. We set $\delta_d = 0.5$ and $\delta_u = 2$. We compute λ_1 , λ_2 and λ_3 as follows: $\lambda_1 = \lambda / [\lambda \sum_{k=1}^{N_i} \sum_{i=1}^{N_w} (S_{r2g}^{k,i} + S_{g2r}^{k,i}) + 2N_r]$, $\lambda_2 = 1 / [\lambda \sum_{k=1}^{N_i} \sum_{i=1}^{N_w} (S_{r2g}^{k,i} + S_{g2r}^{k,i}) + 2N_r]$ and $\lambda_3 = 1/10 [\sum_{j=1}^{N_r} \sum_{l=1}^{N_w} (S_d^{k,j} + S_u^{k,i})]$, where $\lambda = \max(1, 300 - 10 \times (\text{epoch} - 1))$. 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 (Supplementary Fig. 10) and quantitatively evaluate reconstructions with CD and HD metrics, which do not require point-to-point correspondences (Supplementary Table 4). Finally, we visualize the position representations of individual electrode pairs via t-SNE to illustrate the geometrical correlation between different capacitance readouts (Supplementary Fig. 11).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data are publicly available in Edinburgh DataShare with the identifier <https://doi.org/10.7488/ds/3773>⁴⁵.

Code availability

Codes for the implementation of the C2DT are available in Edinburgh DataShare with the identifier <https://doi.org/10.7488/ds/3773>⁴⁵.

References

1. Tuthill, J. C. & Azim, E. Proprioception. *Curr. Biol.* **28**, R194–R203 (2018).
2. Shih, B. et al. Electronic skins and machine learning for intelligent soft robots. *Sci. Robot.* **5**, eaaz9239 (2020).
3. Truby, R. L., Della Santina, C. & Rus, D. Distributed proprioception of 3D configuration in soft, sensorized robots via deep learning. *IEEE Robot. Autom. Lett.* **5**, 3299–3306 (2020).
4. Wang, H., Totaro, M. & Beccai, L. Toward perceptive soft robots: progress and challenges. *Adv. Sci.* **5**, 1800541 (2018).
5. Rus, D. & Tolley, M. T. Design, fabrication and control of soft robots. *Nature* **521**, 467–475 (2015).
6. Cianchetti, M., Laschi, C., Menciassi, A. & Dario, P. Biomedical applications of soft robotics. *Nat. Rev. Mater.* **3**, 143–153 (2018).
7. Cheng, N. et al. Prosthetic jamming terminal device: a case study of untethered soft robotics. *Soft Robot.* **3**, 205–212 (2016).
8. Park, Y. L. et al. Design and control of a bio-inspired soft wearable robotic device for ankle-foot rehabilitation. *Bioinspir. Biomim.* **9**, 016007 (2014).
9. Arnold, T. & Scheutz, M. The tactile ethics of soft robotics: designing wisely for human-robot interaction. *Soft Robot.* **4**, 81–87 (2017).
10. Moin, A. et al. A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition. *Nat. Electron.* **4**, 54–63 (2021).
11. Yu, X. et al. Skin-integrated wireless haptic interfaces for virtual and augmented reality. *Nature* **575**, 473–479 (2019).
12. Thuruthel, T. G., Shih, B., Laschi, C. & Tolley, M. T. Soft robot perception using embedded soft sensors and recurrent neural networks. *Science Robot.* **4**, eaav1488 (2019).
13. To, C., Hellebrekers, T. L. & Park, Y. L. Highly stretchable optical sensors for pressure, strain, and curvature measurement. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2015).
14. Van Meerbeek, I. M., De Sa, C. M. & Shepherd, R. F. Soft optoelectronic sensory foams with proprioception. *Sci. Robot.* **3**, eaau2489 (2018).
15. Kim, T. et al. Heterogeneous sensing in a multifunctional soft sensor for human-robot interfaces. *Sci. Robot.* **5**, eabc6878 (2020).
16. Zhao, Y. et al. Somatosensory actuator based on stretchable conductive photothermally responsive hydrogel. *Sci. Robot.* **6**, eabd5483 (2021).
17. Scharff, R. B. et al. Sensing and reconstruction of 3-D deformation on pneumatic soft robots. *IEEE/ASME Trans. Mechatron.* **26**, 1877–1885 (2021).
18. Glauser, O., Panozzo, D., Hilliges, O. & Sorkine-Hornung, O. Deformation capture via soft and stretchable sensor arrays. *ACM Trans. Graphics* **38**, 1–16 (2019).
19. Leber, A. et al. Soft and stretchable liquid metal transmission lines as distributed probes of multimodal deformations. *Nat. Electron.* **3**, 316–326 (2020).
20. Armanini, C. et al. Flagellate underwater robotics at macroscale: design, modeling, and characterization. *IEEE Trans. Robot.* **38**, 731–747 (2021).
21. Marashdeh, Q. M., Teixeira, F. L. & Fan, L. S. Adaptive electrical capacitance volume tomography. *IEEE Sensors J.* **14**, 1253–1259 (2014).
22. Vaswani, A. et al. Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst.* 5998–6008 (NeurIPS, 2017).
23. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. Annual Conf. North American Chapter of the Association for Computational Linguistics (ACL, 2019)*
24. Dai, Z. et al. Transformer-XL: attentive language models beyond a fixed-length context. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL, 2019)*.
25. Zhao, H., Jia, J. & Koltun, V. Exploring self-attention for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020).
26. Yuan, L. et al. Tokens-to-token ViT: training vision transformers from scratch on ImageNet. In *Proc. IEEE/CVF International Conference on Computer Vision* (IEEE, 2021).
27. Hu, D., Lu, K. & Yang, Y. Image reconstruction for electrical impedance tomography based on spatial invariant feature maps and convolutional neural network. In *Proc. IEEE International Conference on Imaging Systems and Techniques* (IEEE, 2019).
28. Fan, H., Su, H. & Guibas, L. J. A point set generation network for 3D object reconstruction from a single image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017).
29. Wang, R. et al. Real-time soft body 3D proprioception via deep vision-based sensing. *IEEE Robot. Autom. Lett.* **5**, 3382–3389 (2020).
30. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. Araromi, O. A., Rosset, S. & Shea, H. R. High-resolution, large-area fabrication of compliant electrodes via laser ablation for robust, stretchable dielectric elastomer actuators and sensors. *ACS Appl. Mater. Interfaces* **7**, 18046–18053 (2015).

32. Yoon, S. H., Paredes, L., Huo, K. & Ramani, K. MultiSoft: soft sensor enabling real-time multimodal sensing with contact localization and deformation classification. In *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (ACM, 2018).
33. Zhu, Z., Park, H. S. & McAlpine, M. C. 3D printed deformable sensors. *Sci. Adv.* **6**, eaba5575 (2020).
34. Yang, Y., Peng, L. & Jia, J. A novel multi-electrode sensing strategy for electrical capacitance tomography with ultra-low dynamic range. *Flow Meas. Instrum.* **53**, 67–79 (2017).
35. Edelsbrunner, H., Kirkpatrick, D. & Seidel, R. On the shape of a set of points in the plane. *IEEE Trans. Inf. Theory* **29**, 551–559 (1983).
36. Tang, L., Shang, J. & Jiang, X. Multilayered electronic transfer tattoo that can enable the crease amplification effect. *Sci. Adv.* **7**, eabe3778 (2021).
37. Jinkins, K. R. et al. Thermally switchable, crystallizable oil and silicone composite adhesives for skin-interfaced wearable devices. *Sci. Adv.* **8**, eabo0537 (2022).
38. Hang, C. et al. A soft and absorbable temporary epicardial pacing wire. *Adv. Mater.* **33**, 2101447 (2021).
39. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2019).
40. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at arXiv <https://arxiv.org/abs/1412.6980> (2014).
41. Tolgyessy, M., Dekan, M., Chovanec, L. & Hubinsky, P. Evaluation of the azure Kinect and its comparison to Kinect V1 and Kinect V2. *Sensors* **21**, 413 (2021).
42. Geiger, A., Moosmann, F., Car, O. & Schuster, B. Automatic camera and range sensor calibration using a single shot. In *Proc. IEEE International Conference on Robotics and Automation* (IEEE, 2012).
43. Soleimani, V., Mirmehdi, M., Damen, D., Hannuna, S. & Camplani, M. 3D data acquisition and registration using two opposing Kinects. In *Proc. IEEE International Conference on 3D Vision* (IEEE, 2016).
44. Qi, C. R., Yi, L., Su, H. & Guibas, L. J. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proc. Advances in Neural Information Processing Systems* (NeurIPS, 2017).
45. Hu, D., Giorgio-Serchi, F., Zhang, S. & Yang, Y. Stretchable e-skin and transformer enable high-resolution morphological reconstruction for soft robots. *Edinburgh DataShare* <https://doi.org/10.7488/ds/3773> (2022).

Acknowledgements

Y.Y. and F.G.-S. acknowledge the support of the Data Driven Innovation Chancellor's Fellowship at The University of Edinburgh.

D.H. acknowledges the support of the studentship from the School of Engineering, The University of Edinburgh. S.Z. acknowledges the support of the Seed Funding for Strategic Interdisciplinary Research Scheme (SIRS) from The University of Hong Kong and the Germany/Hong Kong Joint Research Scheme (G-HKU707/22) from the Research Grants Council.

Author contributions

Y.Y., F.G.-S., D.H. and S.Z. conceived the concept. Y.Y. and F.G.-S. supervised the project and acquired funding. D.H. and Y.Y. carried out the simulation, 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., F.G.-S. and Y.Y. wrote the manuscript. All authors reviewed and revised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42256-023-00622-8>.

Correspondence and requests for materials should be addressed to Yunjie Yang.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Multi-physics simulation data was generated by COMSOL V5.4 and MATLAB 2021a. Experimental dataset was collected using a 32-channel electrical capacitance tomography system developed by The University of Edinburgh.
Data analysis	Data preprocessing was conducted using MATLAB 2021a and Python V3.7. The training, validation and testing were conducted using Python V3.7 and Pytorch V1.7. Codes are available in Edinburgh DataShare at https://doi.org/10.7488/ds/3773 .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data are publicly available in Edinburgh DataShare with the identifier doi:10.7488/ds/3773 (<https://doi.org/10.7488/ds/3773>).

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

N/A

Population characteristics

N/A

Recruitment

N/A

Ethics oversight

N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The simulation dataset includes 39,334 samples (22,517 for training, 9,721 for validation and 7,096 for testing). The experimental dataset generated by the physical platform includes 30,973 samples (20,693 for training, 6,018 for validation and 4,262 for testing).

Data exclusions

Samples with severe occlusion issues are excluded.

Replication

Deformation reconstruction was successfully implemented on all testing samples.

Randomization

The dataset was divided into training, validation and testing groups randomly.

Blinding

Not applicable.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging