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

Visual-Auditory Proprioception of Soft Finger Shape and Contact

Qinsong Guo*, Ke Yang*, Hanwen Zhao, Haohan Fang, Haoxuan Wang, and Chen Feng✉

***Abstract*— Soft robotic fingers require precise proprioception of both global deformation and local contact to enable safe and dexterous manipulation. Vision-based methods can reconstruct overall shape but struggle under severe occlusion, while audio-only approaches provide complementary cues but lack spatial detail. We present DeepCoFi, a lightweight multimodal proprioception framework that fuses internal camera images with acoustic spectrograms to jointly recover finger geometry and contact. The framework leverages the complementary strengths of vision and acoustics and employs a FoldingNet-based two-stage decoder that first reconstructs global bending and then refines local contact deformations. To support this integration, we introduce a soft finger design that incorporates an exoskeleton-mounted camera and microphone in a single molding step, preserving compliance while enabling multimodal sensing. Experiments on a comprehensive dataset and real-world grasping tasks show that DeepCoFi achieves robust proprioception under occlusion and generalizes effectively to unseen deformations and contact conditions. Open-source resources and project updates are available at ai4ce.github.io/DeepCoFi.**

I. Introduction

Soft robotic fingers combine inherent compliance, virtually unlimited degrees of freedom, and high spatial resolution in physical interaction—capabilities that rigid grippers struggle to match [1]. These advantages have motivated their use in applications ranging from delicate fruit harvesting for precision agriculture [2], to minimally invasive surgical manipulation [3], underwater specimen collection [4], and safe human–robot collaboration in industrial cobots [5]. Reliable manipulation, adaptive grasping, and teleoperation, however, all hinge on accurate sensing of the finger's continuously deforming geometry and external contacts.

Traditional proprioceptive sensing methods for soft fingers generally fall into two categories. Simplified kinematic models, including constant-curvature assumptions and backbone centerline regression [1, 6], are computationally efficient but provide limited detail on surface geometry or contact. Finite element analysis (FEA) simulations [7, 8] achieve precise shape estimation, yet their high computational cost makes them unsuitable for real-time or embedded applications.

With the rise of deep learning, neural networks have increasingly been applied to conventional sensors, such as embedded cameras, to predict soft robot deformation and contact states [9, 10, 11]. For example, Yoo *et al.* employ an internal camera together with a sim-to-real trained model to reconstruct the full three-dimensional shape of a pneumatic soft finger, achieving sub-centimeter accuracy under previously unseen loading conditions [9]. Despite this progress, vision-based methods struggle when bending exceeds the camera's field of view. Severe occlusion hides internal surfaces, preventing the sensors from capturing the true deformation and limiting performance in large-angle, highly occluded scenarios.

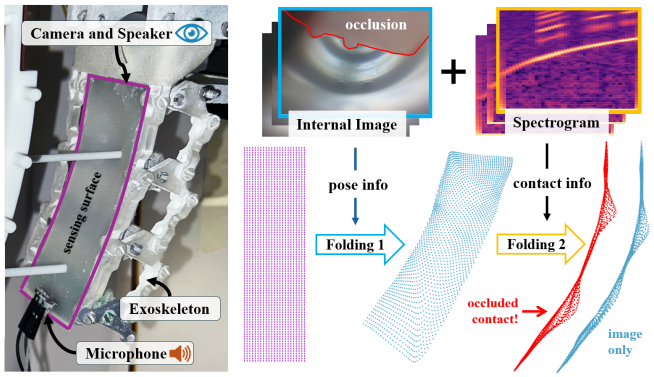

Fig. 1. Overview of our multimodal proprioception framework. A soft robotic finger is instrumented with an internal camera, speaker, and microphone. The camera captures global bending, while spectrograms from acoustic reflections provide complementary contact cues, especially in occluded regions. The modalities are fused and processed through sequential folding modules: **Folding 1** reconstructs the global pose, and **Folding 2** refines the surface with localized contact deformations. The resulting point cloud captures both bending and contact, enabling accurate shape estimation under severe occlusion.

Recent advances in acoustic sensing leverage its immunity to visual occlusion in soft robots. Wall et al. [12] used internal acoustic modulation within pneumatic actuators to infer contact location, while Yoo et al. [13] applied a similar approach to predict deformation. However, these methods either require multiple embedded microphones or are limited to non-bending scenarios, making them unsuitable for highly deformable soft fingers undergoing large bending.

To address these limitations, we propose a lightweight multimodal proprioception system as depicted in Fig. I that combines internal-view camera and acoustic signals to jointly estimate finger shape and contact. Our contributions are:

1) A multimodal proprioception framework that fuses internal visual and acoustic signals to estimate finger shape and contact. This design overcomes the limitations of vision-only methods under severe occlusion and the low spatial resolution of audio-only sensing.
2) A FoldingNet-based neural architecture with two stages: the first reconstructs global bending, and the second refines the surface with localized contact deformations, enabling end-to-end shape and contact prediction.
3) A novel soft finger hardware design that integrates an exoskeleton-mounted camera and microphone within a single molding step, preserving compliance while providing reliable sensing interfaces.

*Equal contribution. ✉Corresponding author.
All authors are with the Center for Robotics and Embodied Intelligence (CREO) at New York University, Brooklyn, NY 11201, USA. {qg2053, ky2276, hz2815, hf2371, hw3061, cfeng}@nyu.edu
This work was supported in part by NSF Grants 2024882 and 2238968.

II. Related Work

A. Proprioceptive Tactile Sensing

Proprioceptive tactile sensing methods can be broadly categorized into electrical, magnetic, optical, and learning-based approaches.

Electrical approaches rely on impedance, capacitance, and the triboelectric effect in conductive materials to measure deformation, typically employing discrete sensing units such as electrodes [14], probes [7], or liquid-metal pockets [15], which inherently limit spatial resolution. Magnetic-based methods instead use Hall effect sensors to detect field changes from embedded elastomer magnets, enabling high-resolution tactile and force measurement with decoupled shear/normal sensing [16, 17].

Optical methods track elastomer surface lighting changes with cameras to compute deformed surface normals [18, 19, 20], achieving higher resolution than magnetic approaches. Neural networks are often integrated to reduce noise [21], though these systems struggle under large deformation, extreme compliance, or occlusion beyond the field of view [10, 11, 13].

Learning-based tactile sensing addresses these limitations by bypassing a priori physical models, making it effective for large-deformation scenarios [10, 22, 23, 24]. Such methods map raw sensor data directly to shape or contact information, enabling high-speed inference—e.g., CNNs using embedded cameras reconstruct point clouds at $\sim$2 ms per frame with $\leq$0.5% error [10]—and zero-shot models predict full 3D configurations from a single internal view with $\sim$8.85 mm Chamfer error [13]. However, reliance on a single sensing modality can still degrade performance under occlusion or outside the sensor field.

B. Auditory Sensing

Vision-based tactile systems often suffer from occlusion when sensing regions lie outside the camera's field of view, motivating auditory sensing—both active and passive—as a complementary modality.

In active auditory sensing, internally or externally generated audio is used to infer hidden shapes and contacts. Stuart modeled the resonance response of sine waves in different-sized elastic chambers, while Shi et al. [25] employed ultrasonic radar arrays to map object height. Learning-based methods further extend active audio for robust manipulation [26, 27] or occlusion recovery [28]. Wall & Brock reconstructed high-resolution contact maps ($\sim$1.67 mm) and achieved 88% braille classification accuracy by analyzing reflections of sweeping sine waves in pneumatic chambers [12]. Later extensions combined passive and active modes, inferring multiple variables (force, material, inflation, temperature) with $\sim$3.7 mm spatial precision.

In passive auditory sensing, no active signal is emitted; instead, naturally occurring sounds are exploited. Liu et al. [29] and Athar et al. [30] analyzed elastomer tapping/grasping acoustics to infer geometry or detect slip. Yoo et al. showed that microphone arrays can predict key shape points and reconstruct meshes with $\sim$4.91 mm Chamfer error [13]. While compact and low-cost, both active and passive systems remain sensitive to background noise and sound propagation, requiring robust signal processing or denoising.

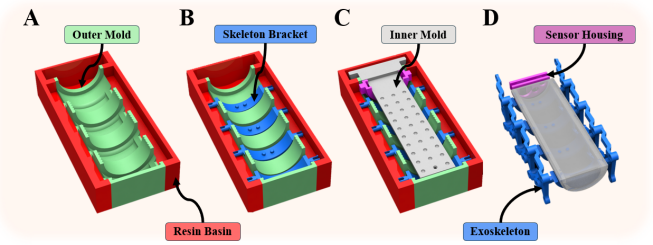

Fig. 2. Fabrication process of the multimodal soft robotic finger. (A) The outer mold and resin basin are prepared. (B) The exoskeleton is inserted and aligned within the mold. (C) The inner mold is placed to define the finger cavity and embed microphone components. (D) The sensor housing is positioned, and the assembly is filled and cured, resulting in a soft finger with integrated exoskeleton braces and sensor mounts.

C. Pose and Contact Estimation

For soft actuators, both pose (large-scale deformation) and contact (localized tactile interaction) are essential for dexterous manipulation, yet a trade-off exists between actuation range and contact resolution.

High-actuation, low-resolution examples include Jin et al.'s triboelectric finger sensor [31], achieving 10 mm resolution only along the finger axis. McCandless et al. [32] demonstrated optical cable sensing for pneumatic actuators, but with tip-only contact detection. Nguyen et al. [11] used a cable-driven continuum actuator with a U-Net and visual markers to reconstruct both contact and shape, though with coarse resolutions (28.86 mm contact, $\sim$5.27° bending).

High-resolution, low-actuation examples include Yoo et al. [13], indirectly inferring contact from shape sensing with 1.92 mm resolution, evaluated via unidirectional Chamfer distance. Guo et al. [33] reconstructed contact under large deformation using mesh-based vision without active actuation, achieving 5.60 mm RMSE. While embedded vision systems yield fine-grained meshes, they suffer occlusion; tactile-only approaches capture high contact resolution but lose global shape fidelity.

In this work, we integrate the fine spatial resolution of vision-based sensing with the occlusion robustness of auditory sensing to achieve simultaneous contact and pose estimation. This multimodal framework enables high-resolution point cloud reconstruction that captures both global shape and local contacts.

III. Methods

A. Problem Formulation

We formulate proprioception of a soft robotic finger as the joint estimation of global bending and local contact from multimodal sensor inputs. The system takes as input an internal camera image $I \in \mathbb{R}^{H\times W\times 3}$ and an acoustic mel-spectrogram $A \in \mathbb{R}^{H\times W\times 3}$. The goal is to reconstruct a point cloud representation of the finger surface capturing both global pose and local contact deformations.

To facilitate learning, we decompose the problem into two stages: bending estimation and contact refinement.

Bending estimation. The bending subnetwork predicts a baseline finger surface without contact

$$h_{\theta_b} : (I, A) \mapsto \hat{P}_b,$$

where $\hat{P}_b$ denotes the estimated bending shape.

Contact estimation. The contact subnetwork refines the predicted bending geometry by incorporating local deformation cues

$$h_{\theta_c} : (\hat{P}_b, I, A) \mapsto \hat{P},$$

producing the final contact-aware point cloud $\hat{P}$.

This formulation decomposes the task into global geometry estimation followed by local deformation refinement, enabling multimodal proprioception of both bending and contact.

B. Hardware Fabrication

The soft robotic finger consists of two primary components: a soft silicone body for sensing and an exoskeleton for actuation. The exoskeleton is fabricated using stereolithography and joined to the silicone body through integrated braces that are co-molded during casting. The complete fabrication process is illustrated in Fig. 2.

During the molding process, the exoskeleton braces and camera housing are positioned in the outer mold, while a microphone is mounted onto the inner mold. The two molds are then assembled and enclosed with a resin basin to prevent overflow. After curing and mold removal, the microphone and exoskeleton braces become embedded within the soft finger. A PiCamera and speaker are subsequently inserted into the camera housing, secured between two spacers.

Finally, the exoskeleton is attached via the integrated braces to complete fabrication.

Compared to pneumatic actuation, exoskeleton-based actuation preserves the flexibility and dexterity of the finger belly, enabling more sensitive force sensing.

C. Data Collection

Due to the presence of the rigid exoskeleton and severe self-occlusion inside the finger cavity, it is infeasible to directly capture complete 3D point clouds representing both bending and localized contact deformation using external depth sensors. We therefore construct ground truth through parametric bending reconstruction followed by depth-consistent deformation injection at predefined grid locations.

Deformation protocol. During acquisition, the finger experiences two deformation modes: bending and contact, while the camera and microphone record synchronized visual and acoustic signals.

Bending. Bending is applied by the servo-driven exoskeleton. To discretize the pose and ensure repeatability under varying contact loads, we use 20 reference pads that conform to the exoskeleton geometry. Each pad corresponds to a specific bending level.

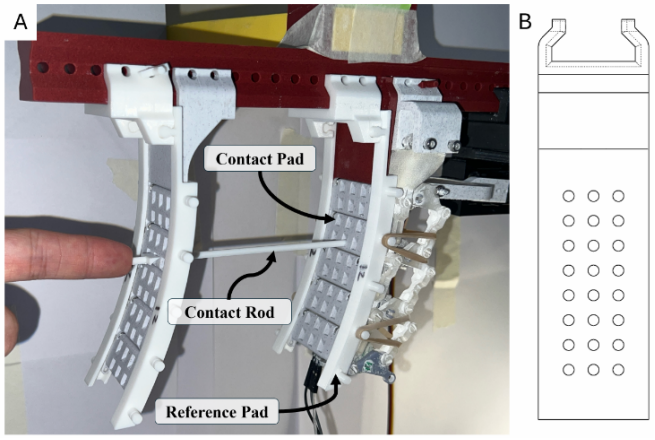

Fig. 3. Contact data collection setup. (A) Experimental configuration with two sets of reference pads and contact pads mounted on the finger. The reference pads ensure repeatable bending geometry, while the contact pads define discrete indentation sites. A thin contact rod is inserted through selected pad slots to apply controlled indentations of known depth on the finger belly. (B) Schematic of the contact pad showing the 3×8 slot grid (24 sites) used to parameterize possible contact locations.

Contact. One or more thin rods are inserted through contact pads integrated into the reference-pad assembly, producing indentations at prescribed locations on the finger belly (Fig. 3). Each pad contains an 8×3 slot grid, yielding 24 possible sites per bending level. Both single- and multi-contact configurations are realized by selecting one or more sites from this grid.

Sensing and preprocessing.

Image. Internal images are captured by a PiCamera 2 at 1080p, then center-cropped and resized to 256×256.

Audio. A sine sweep from 1 kHz to 4 kHz is emitted at 4 Hz, and the reflected signal inside the finger cavity is recorded by the embedded microphone. From this signal we compute a normalized 256×256 mel-spectrogram, following prior acoustic proprioception practice [12].

Bending ground truth. We capture the externally visible surface geometry of the bent finger using a pair of Azure Kinect depth cameras and fuse the measurements to obtain point clouds under bending-only conditions. The belly centerline across bending levels is fitted with a parabolic curve $y \approx ax^2$, where the curvature parameter $a \in [0.00, 6.92]$ characterizes the global bending state. Using this fitted curvature, a smooth 3D bending surface is reconstructed by sweeping a 2D cross-section along the parametric path, yielding a continuous representation of the finger belly geometry.

Contact supervision. Following the grid-based contact setup described in Fig. 3, we reconstruct contact deformation according to the same 8×3 belly discretization. Gaussian perturbations are applied at the corresponding grid locations on the reconstructed bending surface, with amplitudes and widths calibrated to match the empirically measured indentation depths (3 mm and 7 mm). The resulting reconstructed bending belly with contact deformation serves as the final ground truth for model training.

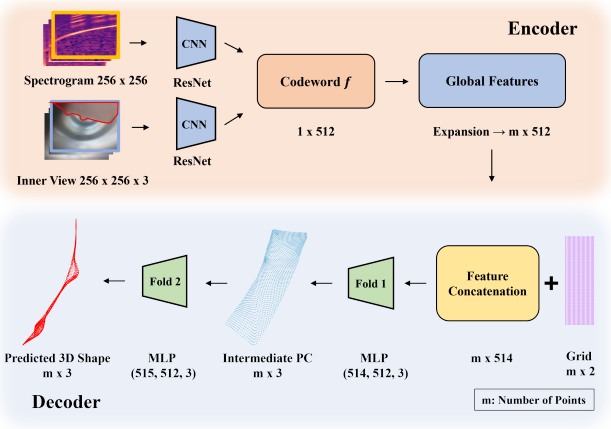

Fig. 4. DeepCoFi model architecture. The proposed framework encodes multimodal proprioceptive inputs—internal images and spectrograms—through ResNet-18 backbones, producing a fused latent codeword $f \in \mathbb{R}^{d_f}$. This codeword is expanded and concatenated with a 2D prototype grid of m points. The decoder then applies two sequential folding modules: Fold 1 reconstructs the global bending shape, and Fold 2 refines local contact deformations. The final output is a predicted point cloud $\hat{P} \in \mathbb{R}^{m\times 3}$ that captures both overall pose and contact.

D. Model Architecture

Visual and audio encoders. The multimodal inputs consist of the internal image $I \in \mathbb{R}^{H\times W\times 3}$ and the mel-spectrogram $A \in \mathbb{R}^{H\times W\times 3}$. Both are processed by convolutional encoders based on ResNet-18. Each encoder extracts a feature embedding of dimension d_e, which is projected to d_p through a linear layer. The visual and audio embeddings are concatenated to form a fused latent representation

$$f \in \mathbb{R}^{d_f}, \quad d_f = 2d_p.$$

FoldingNet-based decoder. To reconstruct the finger geometry, we employ a FoldingNet-style decoder [34]. The decoder begins with a two-dimensional prototype grid of m points sampled from a uniform mesh. The fused codeword f is concatenated with the prototype coordinates and passed through a sequence of folding modules, each implemented as a multilayer perceptron (MLP) that progressively deforms the prototype into a point cloud. Two folding stages are applied in sequence: the first captures global bending, and the second refines local contact deformations. The overall pipeline is shown in Fig. 4, and the model output is the reconstructed point cloud.

$$\hat{P} = h_\theta(I, A) \in \mathbb{R}^{m\times 3}.$$

DeepSoRo-inspired baseline. As a comparison, we design an alternative decoder inspired by DeepSoRo [10]. This baseline separates bending and contact reasoning. The bending shape $\hat{P}_b$ is first predicted by a pretrained module and encoded by PointNet into a global feature of dimension d_g. This geometric descriptor is concatenated with the visual and audio embeddings, producing a multimodal latent vector in $\mathbb{R}^{d_g+d_f}$. A stack of fully connected layers then regresses per-point contact offsets, which are added to the coordinates of $\hat{P}_b$ to yield the final estimate $\hat{P}$.

Loss functions. The networks are trained on simulated point clouds $P \in \mathbb{R}^{m\times 3}$ derived from the parameterized finger model. The objective is a weighted sum of the Chamfer Distance (CD) and an asymmetric Hausdorff Distance (AHD), which emphasizes accuracy along the y-axis to better capture contact.

Chamfer Distance:

$$\mathcal{L}_{\text{Chamfer}} = \frac{1}{|\hat{P}|}\sum_{\hat{p}\in\hat{P}}\min_{p\in P}\|\hat{p}-p\|_2^2 + \frac{1}{|P|}\sum_{p\in P}\min_{\hat{p}\in\hat{P}}\|p-\hat{p}\|_2^2 \quad (1)$$

Asymmetric Hausdorff Distance:

$$\mathcal{L}_{\text{Hausdorff}} = \frac{1}{B}\sum_{i=1}^{B}\max_{\hat{p}\in\hat{P}_i}\min_{p\in P_i}\left\|W(\hat{p}-p)\right\|_2, \quad (2)$$

where $W = \text{diag}(1, w_y, 1)$ increases the penalty in the y-direction.

Total loss:

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{Chamfer}} + (1-\lambda)\mathcal{L}_{\text{Hausdorff}}, \quad (3)$$

with $\lambda \in [0, 1]$ and w_y as tunable hyperparameters.

E. Grasping Experiment

Setup. To evaluate the proposed proprioception model in a functional task, we conducted object grasping experiments using two soft robotic fingers of the design described above. The fingers were mounted face-to-face with a gap sufficient to accommodate test objects, as shown in Fig. 5.

Independent servo actuation controlled the bending angle and grasping force of each finger. During each trial, the actuators executed convergent motions to emulate natural grasping. To reflect real-world variability, actuator forces were randomly perturbed, and the vertical position of the object was slightly shifted across trials.

PointNet-based classification. For evaluation, the fused multimodal inputs (I, A) were first processed by the best-performing architecture from Sec. III, producing a predicted point cloud $\hat{P} \in \mathbb{R}^{m\times 3}$. This point cloud was then encoded by PointNet [35], with the feature-transform module disabled and dropout rate set to 0.5. The pooled global feature of dimension (B, d_g), where $d_g = 1024$, was passed through two fully connected layers with batch normalization and ReLU activation, yielding $(B, 256)$. A final linear classifier mapped

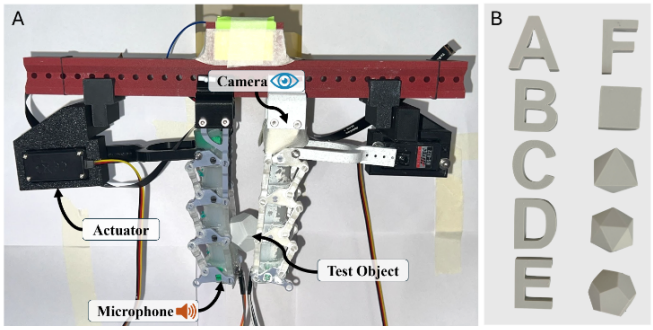

Fig. 5. Grasping experiment. (A) Experimental setup with two actuated soft robotic fingers positioned face-to-face. Each finger is equipped with an internal camera, speaker, and microphone, and is driven by an independent servo actuator. A test object is placed between the fingers while the actuators perform convergent motions to emulate grasping. (B) Set of test objects used for classification, including letter A-F, cube, octahedron, icosahedron, and dodecahedron.

TABLE I

KEY SYMBOLS AND HYPERPARAMETERS.

Symbol	Meaning	Value
I	Internal RGB image	$256 \times 256 \times 3$
A	Mel-spectrogram	256×256
m	Prototype grid points	2025 (45×45)
d_e	Encoder feature dim	512
d_p	Projected feature dim	256
d_f	Fused feature dim	512 $(2d_p)$
d_g	PointNet global feature	1024
f	Multimodal latent codeword	$\mathbb{R}^{512}$
$P, \hat{P}$	GT / predicted point cloud	2025×3
B	Batch size	$-$ (varies)
C	Object classes (grasping)	10
λ	Chamfer vs. Hausdorff weight	$[0, 1]$
w_y	Hausdorff y-axis weight	≥ 1 (tuned)

this representation to $C = 10$ output logits, corresponding to the object classes.

The network was trained using cross-entropy loss on the logits $z \in \mathbb{R}^{B \times C}$. At inference, the predicted object label was obtained as $\arg\max(z)$. The notation and hyperparameters used throughout the model are summarized in Table I.

IV. EXPERIMENTS

A. Dataset

The dataset used to evaluate the proposed architecture contains 1,617 unique contact–bending conditions. It includes 980 single-contact samples (24 locations at 3 mm depth, 24 locations at 7 mm depth, and one no-contact case for each of 20 bending angles), 563 two-contact samples with randomly generated pairs, and 74 three-contact samples.

Bending angles 5 and 16 are randomly selected and reserved for testing, representing mild and large curvature regimes within the overall bending range. The remaining samples are split into training and validation sets with an 8:2 ratio.

B. Implementation

Models are trained using Adam with a learning rate of 1×10^{-4} under a three-stage training regime:

Pretraining. FoldingNet is trained on single-contact data for 500 epochs with $w_y = 1$ and $\lambda = 1$, using only Chamfer distance to encourage uniform point distribution.

Multi-contact fine-tuning. The model is fine-tuned on two- and three-contact samples for 500 epochs with $w_y = 2$ and $\lambda = 0.5$. An asymmetric Hausdorff loss is introduced to emphasize deformations along the y-axis, improving sensitivity to indentation depth and location while preventing degenerate bending–contact tradeoffs.

Final fine-tuning. The network is further trained on the full dataset for 500 epochs with $w_y = 1$ and $\lambda = 0.5$. A symmetric Hausdorff loss is reinstated to encourage balanced reconstruction across bending and contact conditions.

The prototype point cloud is initialized as a 45×45 grid uniformly spanning $(-1, -1)$ to $(1, 1)$. All subnetworks are updated jointly; no weights are frozen at any stage.

C. Network Variations

Epochs trained. Early stopping is applied when the validation loss begins to increase, typically before reaching the full 1,500 epochs. Performance is reported at this stopping point. *A2 (TwoFoldsShort)* shares the same architecture as *A1 (TwoFoldsControl)*, but each stage is limited to 250 epochs.

Number of folding layers.

- *B1 (OneBigFold)*: replaces the two folding modules with a single, deeper module consisting of five hidden layers $(514, 512, 1024, 1024, 512, 3)$.
- *B2 (ThreeFolds)*: retains the original folding subnetwork but increases the depth to three folding modules. Each stage is trained for 700 epochs before early stopping.
- *B3 (OneSmallFold)*: uses only one folding module from the original two-stage design.

D. Ablation Studies

To assess the contribution of each modality, we evaluate performance under unimodal input:

C1 (audio only): the image modality is removed; audio features are duplicated to match the fusion dimension.

C2 (image only): the audio modality is removed; image features are duplicated accordingly.

E. Baselines

We compare the proposed method against three representative baselines:

- **MLP baseline.** A simple multi-layer perceptron that directly maps fused image and acoustic features to contact predictions. It consists of several fully connected layers with ReLU activations and dropout, without explicitly modeling the spatial structure of deformations. This baseline tests whether global multimodal statistics alone are sufficient for contact inference.
- **DeepSoRo-MLP.** A variant inspired by the DeepSoRo [10]. Acoustic spectrogram and image features are fused with a reconstructed bending shape, and the concatenated representation is decoded by an MLP to predict contact-related deformation. This mirrors the two-stage sensing pipeline in [10], where sensory inputs are embedded into a latent space and decoded into shape.
- **DeepSoRo-PointNet.** Another DeepSoRo-based variant that incorporates PointNet [35] into the pipeline. The reconstructed bending shape is first encoded by PointNet to extract point-wise structural features, which are then fused with multimodal embeddings before contact prediction. This design draws on the use of FoldingNet and point-based decoders in [10], highlighting the advantage of point cloud representations.

All baselines are trained with the same optimizer (Adam, learning rate 1×10^{-4}), three-stage training curriculum (single-contact pretraining, multi-contact fine-tuning, and full-dataset fine-tuning), and combined Chamfer–Hausdorff loss. This ensures that differences in performance arise from architectural design rather than optimization settings.

TABLE II

QUANTITATIVE RESULTS ON TEST (UNSEEN BENDING CONDITIONS) AND VALIDATION SETS. DISTANCES ARE REPORTED IN MILLIMETERS.

Category	Index	Input Features	Encoder	Decoder	Layers	Layer Depth	Epochs Trained	Test Set		Validation Set	
								CD↓(mm)	AHD↓(mm)	CD↓(mm)	AHD↓(mm)
Training Epochs	A1: TwoFoldsControl	Image + Audio	/	FoldingNet	2	3	1140	4.23	9.60	1.51	3.88
	A2: TwoFoldsShort	Image + Audio	/	FoldingNet	2	3	750	4.83	11.13	1.79	4.64
Number of Folds	B1: OneBigFold	Image + Audio	/	FoldingNet	1	5*	1210	5.22	11.81	1.52	4.68
	B2: ThreeFolds	Image + Audio	/	FoldingNet	3	3	2100	5.19	13.12	1.24	2.46
	B3: OneSmallFold	Image + Audio	/	FoldingNet	1	3	1190	6.11	12.80	1.63	4.24
Ablation	C1: Audio_only	Audio	/	FoldingNet	2	3	1500	4.47	10.66	2.98	7.33
	C2: Image_only	Image	/	FoldingNet	2	3	1500	3.55	9.44	1.50	3.64
DeepSoRo	D1: DeepsoroControl	Image + Audio	/	MLP	/	/	1130	2.42†	7.67†	1.83†	7.31†
	D2: DeepsoroPointNet	Image + Audio	PointNet	PointNet+MLP	/	/	1210	2.44†	8.34†	1.74†	7.26†
MLP	E1: pure_MLP	Image + Audio	/	MLP	/	/	1330	2.51†	7.74†	1.81†	7.16†

* the widest hidden layer is of $m \times 1024$. ↓ lower the better. † completely failed to estimate contact.

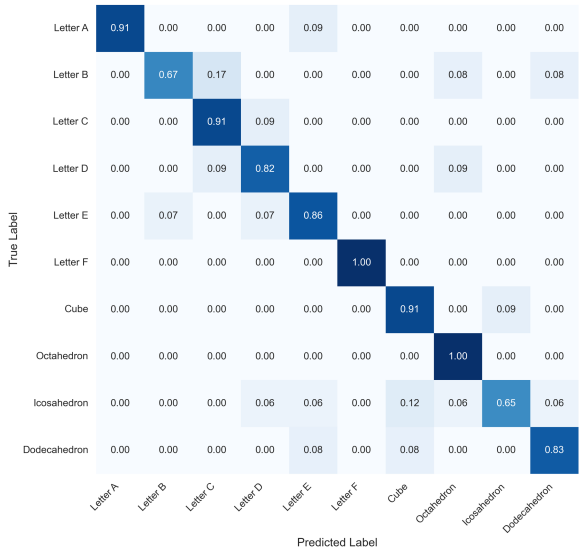

Fig. 6. Confusion matrix for object classification. An 85% overall accuracy is achieved on the test set. The Cube and Octahedron achieve near-perfect recall, while the geometrically complex Icosahedron is the most difficult. All letters are recognized robustly except for B, which is visually similar to C.

F. Performance

Quantitative results. We evaluate all models using Chamfer Distance (CD) and Asymmetric Hausdorff Distance (AHD, $w_y = 2$), reported on the test and validation sets. The test set measures interpolation to unseen bending angles, while the validation set measures contact prediction under seen bending. Percentage changes are computed relative to the best-performing baseline, *D1 (DeepSoRo-MLP)*. The full quantitative results are summarized in Table II.

On the test set, all FoldingNet variants underperform compared with *D1*, indicating weaker generalization to unseen bending angles. Among the multimodal variants, *A1 (TwoFoldsControl)* achieves the best results, outperforming deeper or wider models (*B1, B2*) by an average of 40% in CD and 37.5% in AHD. This suggests that increasing folding depth or width does not improve interpolation. Smaller models (*B3*) and shorter training regimes (*A2*) also fail to reach comparable accuracy, underscoring the need for an appropriate balance of model capacity and training duration.

In contrast, on the validation set all multimodal FoldingNet variants surpass the baseline. *B2 (ThreeFolds)* achieves the largest error reduction (32% in CD and 66% in AHD), followed by *A1* (18% in CD and 47% in AHD). The consistent improvements in AHD highlight the advantage of FoldingNet architectures for contact reconstruction, where DeepSoRo-based decoders struggle. Interestingly, the image-only ablation (*C2*) outperforms other FoldingNet variants despite lacking audio input. This effect can be attributed to stronger bending reconstruction at the cost of weaker contact inference, a trade-off we examine in more detail below.

Qualitative results. Representative model variations are compared across five challenging scenarios involving multi-contact and large bending conditions, as illustrated in Fig. 7.

The DeepSoRo baseline (*D1*) reconstructed bending shapes with high fidelity but consistently failed to localize contacts, even when they were not occluded. This highlights the strength of FoldingNet-based decoders in inferring contact, even when their quantitative metrics appear lower.

Among FoldingNet variants, *A1 (TwoFoldsControl)* consistently produced the most accurate contact localization. For example, while *B1 (OneBigFold)* generated slightly finer contact profiles in Condition 3, it missed the third occluded contact in Condition 2. In Condition 4, all models deviated from the ground truth bending shape, yet *A1 (TwoFoldsControl)* predicted the contact position and profile most accurately, whereas *B2 (ThreeFolds)*, despite stronger quantitative metrics, mislocalized the contact.

A common failure mode of FoldingNet models occurs when two occluded contacts are close together, leading to an averaged indentation. In Condition 5, *A1 (TwoFoldsControl)* predicted a single broader indentation between two ground truth sites, while both ablation models omitted the distant one and *C2 (image-only)* mislocalized the nearer site.

The importance of audio is especially evident in Conditions 3 and 4. The image-only ablation (*C2*) failed to reconstruct the occluded contact in Condition 3 and missed the second contact at the unseen bending angle of five in Condition 4. In contrast, all multimodal FoldingNet variants, including *A1 (TwoFoldsControl)*, *B1 (OneBigFold)*, and *B2 (ThreeFolds)*, successfully reconstructed both contacts and maintained accurate bending. This demonstrates the role of audio in overcoming occlusion and enabling robust generalization to unseen scenarios.

G. Real-World Validation

Grasping objects. Ten 3D-printed objects are used in the grasping validation experiment. Six are English letters (A

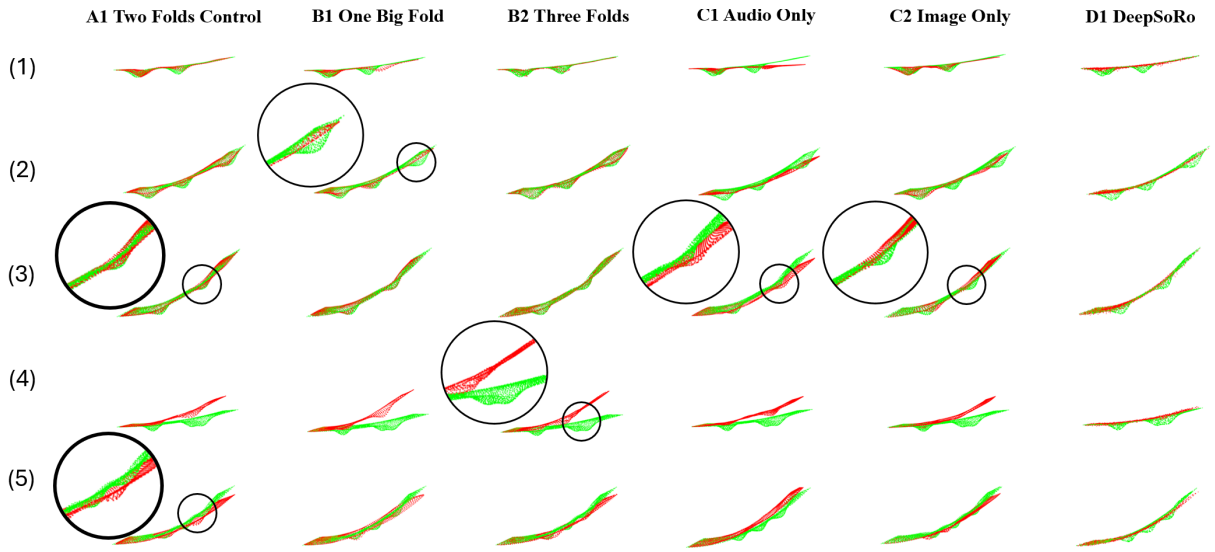

Fig. 7. Qualitative results for five representative conditions. Rows 1–3 correspond to validation set conditions, while Rows 4–5 are from the test set. Each column displays the output of a different network variation, with ground truth shown in green and predictions in red. *A1 (TwoFoldsControl)* consistently localizes contacts, even at occluded sites under large bending, whereas the image-only *C2* struggles. The DeepSoRo baseline fails to reconstruct contacts across all conditions, while variants with stronger quantitative metrics either omit (*B1*) or misplace (*B2*) distant contacts.

to F), and four are polyhedral shapes: cube, octahedron, icosahedron, and dodecahedron. All objects were designed with comparable widths to fit the fixed gap between the two fingers in the setup.

Experiment protocol. Each object is tested in 60 trials, resulting in 600 samples in total. In each trial, the object is placed near the lower portion of the fingers, where visual occlusion is more pronounced. Actuation forces are then applied to induce grasping deformations, with slight random variations introduced across trials to improve robustness and variability. Synchronized audio and visual data from both fingers are collected to train an object classifier.

Results. The grasping experiment validates DeepCoFi's ability to discriminate among objects with both alphabetic and geometric shapes. The best-performing model, selected at epoch 736 based on validation loss, achieved an overall classification accuracy of 85%. As shown in Fig. 6, the confusion matrix reveals near-perfect recall for the cube and octahedron, while the icosahedron was the most challenging, occasionally misclassified as other polyhedra. Among letters, F and C were recognized robustly, whereas visually similar outlines such as B versus C or C versus D led to higher error rates. Despite these ambiguities, per-class recalls exceeded 80% for most categories, demonstrating that the reconstructed point-cloud features provide strong discriminative power across diverse object types.

V. Conclusion

DeepCoFi is a multimodal proprioception framework that accurately reconstructs both global pose and localized contacts of a soft robotic finger, maintaining robust performance under severe occlusion and large bending. It achieves an average Chamfer distance error of 1.51 mm on the validation set, representing a notable improvement in sensing resolution over prior works on contact prediction for highly actuated soft robotic fingers. Contact sensing accuracy, quantified by asymmetric Hausdorff distance, improves by 47% compared to the DeepSoRo-inspired baseline. Furthermore, the reconstructed point clouds obtained from grasping ten objects are used to train a classifier, which achieves 85% accuracy.

Limitations. This work has several limitations. First, due to similar resultant deformations, bending and large-contact predictions are often coupled. Second, for closely spaced occluded contacts, the model tends to produce an averaged indentation, reducing effective resolution. Third, robustness under noisy interference has not been assessed, posing potential risks in high-noise environments.

Future Work. We will explore different microphone arrays and waveforms to mitigate noise and enhance contact localization, and assess the generalizability of the DeepCoFi multimodal platform across diverse soft robotic systems.

Acknowledgment

This work was supported by NYU IT High Performance Computing resources, services, and staff expertise. The authors acknowledge the use of generative AI tools for grammar polishing, clarity improvement, and text shortening during manuscript preparation. All technical content, including the problem formulation, model design, experiments, and conclusions, was developed by the authors. Chen Feng holds concurrent appointments as an NYU Professor and as an Amazon Scholar. This paper describes work performed at NYU and is not associated with Amazon.

References

[1] Z. Chen, F. Renda, A. Le Gall, L. Mocellin, M. Bernabei, T. Dangel, G. Ciuti, M. Cianchetti, and C. Stefanini, "Data-Driven Methods Applied to Soft Robot Modeling and Control: A Review," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 2241–2256, 2025. 1

[2] A. Qiu, C. Young, A. Gunderman, M. Azizkhani, Y. Chen, and A.-P. Hu, "Tendon-Driven Soft Robotic Gripper with

Integrated Ripeness Sensing for Blackberry Harvesting," Feb. 2023. 1

[3] M. Cianchetti, T. Ranzani, G. Gerboni, I. De Falco, C. Laschi, and A. Menciassi, "STIFF-FLOP surgical manipulator: Mechanical design and experimental characterization of the single module," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2013, pp. 3576–3581. 1

[4] N. R. Sinatra, C. B. Teeple, D. M. Vogt, K. K. Parker, D. F. Gruber, and R. J. Wood, "Ultragentle manipulation of delicate structures using a soft robotic gripper," *Science Robotics*, vol. 4, no. 33, p. eaax5425, Aug. 2019. 1

[5] C. Firth, K. Dunn, M. H. Haeusler, and Y. Sun, "Anthropomorphic soft robotic end-effector for use with collaborative robots in the construction industry," *Automation in Construction*, vol. 138, p. 104218, June 2022. 1

[6] A. Melingui, J. J.-B. Mvogo Ahanda, O. Lakhal, J. B. Mbede, and R. Merzouki, "Adaptive Algorithms for Performance Improvement of a Class of Continuum Manipulators," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 9, pp. 1531–1541, Sept. 2018. 1

[7] L. Jin, X. Zhai, W. Xue, K. Zhang, J. Jiang, M. Bodaghi, and W.-H. Liao, "Finite element analysis, machine learning, and digital twins for soft robots: State-of-arts and perspectives," *Smart Materials and Structures*, vol. 34, no. 3, p. 033002, Mar. 2025. 1, 2

[8] P. Jindal, F. Worcester, F. L. Siena, C. Forbes, M. Juneja, and P. Breedon, "Mechanical behaviour of 3D printed vs thermoformed clear dental aligner materials under non-linear compressive loading using FEM," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 112, p. 104045, Dec. 2020. 1

[9] U. Yoo, H. Zhao, A. Altamirano, W. Yuan, and C. Feng, "Toward Zero-Shot Sim-to-Real Transfer Learning for Pneumatic Soft Robot 3D Proprioceptive Sensing," Mar. 2023. 1

[10] 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. 1, 2, 4, 5

[11] T. Nguyen, Q. Luu, D. Nguyen, and V. Ho, "ConTac: Continuum-Emulated Soft Skinned Arm with Vision-based Shape Sensing and Contact-aware Manipulation," in *Robotics: Science and Systems XX*. Robotics: Science and Systems Foundation, July 2024. 1, 2

[12] V. Wall and O. Brock, "A Virtual 2D Tactile Array for Soft Actuators Using Acoustic Sensing," Aug. 2022. 1, 2, 3

[13] U. Yoo, Z. Lopez, J. Ichnowski, and J. Oh, "POE: Acoustic Soft Robotic Proprioception for Omnidirectional End-effectors," Jan. 2024. 1, 2

[14] N. Wettels, V. J. Santos, R. S. Johansson, and G. E. Loeb, "Biomimetic Tactile Sensor Array," *Advanced Robotics*, 2008. 2

[15] J. Xu, Z. Xie, H. Yue, Y. Lu, and F. Yang, "A triboelectric multifunctional sensor based on the controlled buckling structure for motion monitoring and bionic tactile of soft robots," *Nano Energy*, vol. 104, p. 107845, 2022. 2

[16] Y. Yan, Z. Hu, Z. Yang, W. Yuan, C. Song, J. Pan, and Y. Shen, "Soft magnetic skin for super-resolution tactile sensing with force self-decoupling," *Science Robotics*, vol. 6, no. 51, p. eabc8801, 2021. 2

[17] R. Bhirangi, V. Pattabiraman, E. Erciyes, Y. Cao, T. Hellebrekers, and L. Pinto, "AnySkin: Plug-and-play Skin Sensing for Robotic Touch," 2024. 2

[18] W. Yuan, S. Dong, and E. Adelson, "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force," *Sensors*, vol. 17, no. 12, p. 2762, 2017. 2

[19] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The TacTip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 2018. 2

[20] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor With Application to In-Hand Manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020. 2

[21] J. A. Eyzaguirre, M. Oller, and N. Fazeli, "Tactile Neural De-rendering," 2024. 2

[22] U. Yoo, J. Francis, J. Oh, and J. Ichnowski, "KineSoft: Learning Proprioceptive Manipulation Policies with Soft Robot Hands," May 2025. 2

[23] A. Zhang, R. L. Truby, L. Chin, S. Li, and D. Rus, "Vision-Based Sensing for Electrically-Driven Soft Actuators," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 509–11 516, 2022. 2

[24] N. F. Lepora, Y. Lin, B. Money-Coomes, and J. Lloyd, "DigiTac: A DIGIT-TacTip Hybrid Tactile Sensor for Comparing Low-Cost High-Resolution Robot Touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9382–9388, 2022. 2

[25] Q. Shi, Z. Sun, X. Le, J. Xie, and C. Lee, "Soft Robotic Perception System with Ultrasonic Auto-Positioning and Multimodal Sensory Intelligence," *ACS Nano*, 2023. 2

[26] J. Jones, O. Mees, C. Sferrazza, K. Stachowicz, P. Abbeel, and S. Levine, "Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, USA, 2025. 2

[27] K. Zhang, D.-G. Kim, E. T. Chang, H.-H. Liang, Z. He, K. Lampo, P. Wu, I. Kymissis, and M. Ciocarlie, "Vibecheck: Using active acoustic tactile sensing for contact-rich manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2504.15535 2

[28] X. Yi, J. Lee, and N. Fazeli, "Visual-auditory Extrinsic Contact Estimation," 2025. 2

[29] J. Liu and B. Chen, "SonicSense: Object Perception from In-Hand Acoustic Vibration," 2024. 2

[30] S. Athar, X. Zhang, J. Ueda, Y. Zhao, and Y. She, "VibTac: A High-Resolution High-Bandwidth Tactile Sensing Finger for Multi-Modal Perception in Robotic Manipulation," *IEEE Transactions on Haptics*, pp. 1–12, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10965524 2

[31] T. Jin, Z. Sun, L. Li, Q. Zhang, M. Zhu, Z. Zhang, G. Yuan, T. Chen, Y. Tian, X. Hou, and C. Lee, "Triboelectric nanogenerator sensors for soft robotics aiming at digital twin applications," *Nature Communications*, vol. 11, no. 1, p. 5381, Oct. 2020, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-020-19059-3 2

[32] M. McCandless, F. J. Wise, and S. Russo, "A Soft Robot with Three Dimensional Shape Sensing and Contact Recognition Multi-Modal Sensing via Tunable Soft Optical Sensors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, United Kingdom: IEEE, 2023, pp. 573–580. 2

[33] N. Guo, X. Han, S. Zhong, Z. Zhou, J. Lin, F. Wan, and C. Song, "Reconstructing Soft Robotic Touch via In-Finger Vision," *Advanced Intelligent Systems*, vol. 6, no. 10, p. 2400022, 2024. 2

[34] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," 2018. [Online]. Available: https://arxiv.org/abs/1712.07262 4

[35] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4, 5