Topology-Agnostic Mesh Reconstruction of Deformable Objects from Sparse Touch

Everest Yang
Brown University
everest_yang@brown.edu

***Abstract*—Estimating the full shape of a deformable object is especially challenging when vision is unavailable: in the dark, inside an opaque bag, behind the manipulating hand, or under heavy self-occlusion. Touch is the natural sensor in these settings, but touches are sparse and local. We present a single *topology-agnostic* estimator that reconstructs the full mesh of a deformable object from only a few touches and no vision, using one permutation-invariant cross-attention architecture that handles a 1D rope, a 2D cloth, and a 3D volumetric soft body. The learned estimator reduces reconstruction error by roughly two-thirds relative to non-learned geometric mesh completion and a Gaussian-process surface baseline, and it outperforms a simpler global-pool set encoder, with the gap growing as more touches are observed. We then show that the estimator's deep-ensemble uncertainty can be used to learn where to touch next, which lowers error further and beats both random touching and a Gaussian-process active baseline at sparse budgets. This gain is modest on average but grows with self-occlusion and on the error tail. When vision is also available, where to touch barely matters, motivating the vision-free setting we study.**

I. Introduction

Manipulating a deformable object, whether cloth, rope, or cable, starts with a hard perception problem: its configuration is effectively infinite-dimensional, so a robot must estimate the object's current shape before it can fold, route, or grasp it. State estimation is thus a prerequisite for almost every deformable-manipulation task.

Almost all existing deformable perception assumes vision [1, 3, 13]. Yet the settings where deformable manipulation matters most are often those where vision is degraded or unavailable. Consider reaching into a bag or drawer, operating in darkness, manipulating an object in-hand where the gripper hides it, or handling heavily crumpled cloth whose folded-under layers no external camera can see. Touch is then the natural and often the only reliable sensor. But each contact is local and costly, so a practical system can afford only a handful of them. This raises two coupled questions. First, can a robot reconstruct the *full* shape from a handful of sparse, noisy touches and no vision? Second, can it decide *where to touch next* to learn the most per touch? We study both and separate them: estimation first, active sensing on top.

Our **primary contribution** is a single *topology-agnostic* estimator that reconstructs the full vertex mesh from sparse touch alone. It is a permutation-invariant cross-attention network [15, 17]: touch observations are decoded via a fixed bank of grid-position queries, one per output vertex, so changing only the query-grid dimensionality yields the same model for a 1D rope, 2D cloth, or 3D soft body. Prior active-touch work targets rigid objects [11] or vision-led cable splines [8]; we are not aware of a tactile-only full-mesh estimator demonstrated across all three. A deep ensemble [4] provides per-vertex uncertainty.

Our **secondary contribution** is learned active sensing: an acquisition policy trained against an expected-error-reduction oracle uses ensemble uncertainty to pick the next touch. It beats random touching and a Gaussian-process active baseline [16] at sparse budgets, though the average gain is modest and grows with self-occlusion and on the error tail. When vision is also available, vision and the learned prior together pin down the hidden state and touch placement barely matters, motivating the vision-free setting we study.

In summary, we contribute: (i) a topology-agnostic cross-attention estimator for full-mesh reconstruction from sparse touch across 1D, 2D, and 3D deformables; (ii) a learned active policy with occlusion-dependent gains at sparse budgets; and (iii) a negative result delimiting when active placement helps.

II. Related Work

Deformable estimation from vision: Most deformable shape estimation recovers full state passively from vision: garment-mesh completion from depth [1], cloth-mesh reconstruction with occlusion reasoning [3], diffusion models that generate full cloth states for model-predictive control [13], and rope state from a single occluded point cloud [6]. Separate lines of work learn dynamics across topologies [7, 5], but assume near-full visual state and predict dynamics rather than reconstructing a mesh from sparse touch. These methods are vision-led and passive. We instead reconstruct from touch alone, actively select observations, and use one model across 1D, 2D, and 3D objects.

Active tactile shape estimation: Choosing where to touch to recover shape has been studied mainly for *rigid* objects: learned next-best-touch from vision and touch on CAD models [11], classical GP uncertainty sampling [16, 2], and active visuo-haptic completion that contacts points of greatest ensemble disagreement [10], in the spirit of expected-error reduction [9]. Visuo-tactile methods such as NeuralFeels [12] reconstruct in-hand shape passively with vision. On deformables, Mazza *et al.* [8] probe visually occluded cable segments, but are vision-led, use a geometric heuristic rather than a learned policy and output a 1D spline rather than a mesh. Act-VH's disagreement heuristic is the closest acquisition precedent; we

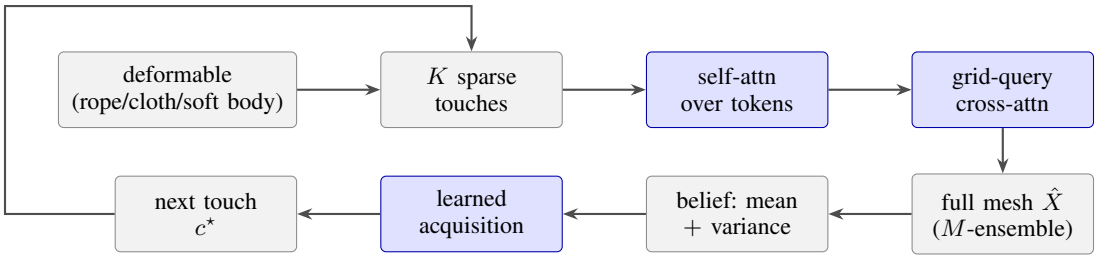

Fig. 1. One topology-agnostic estimator reconstructs a 1D, 2D, or 3D deformable mesh from sparse touches via grid-query cross-attention. A deep ensemble gives a per-vertex belief and a learned acquisition picks the next touch (blue: learned).

show the analogous variance rule degrades below random on deformables, motivating a learned policy.

III. Method

Problem: A deformable object is a mesh of N vertices on a regular grid $\mathcal{G}$ of shape (H, W) for a sheet, $(H, 1)$ for a linear object, or (D, H, W) for a volume, with configuration $X \in \mathbb{R}^{N\times 3}$ drawn from a distribution over settled configurations. A touch at vertex i returns a noisy position $o_i = x_i + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Given observations O_S over touched vertices S, the estimator $\hat{X} = f(O_S)$ reconstructs the full mesh. We measure the mean per-vertex error $\mathcal{E}(S) = \frac{1}{N}\sum_i \|\hat{x}_i(O_S) - x_i\|_2$.

Topology-agnostic estimator: Figure 1 summarizes the pipeline. Each observation is encoded as a token concatenating its noisy 3D coordinates with the normalized grid index of the touched vertex. The grid index has $\dim(\mathcal{G}) \in \{1, 2, 3\}$ components which is the only thing that changes across topologies. Tokens are embedded by an MLP to width d=128 and mixed by one multi-head self-attention layer so observations contextualize one another. A fixed bank of *query* tokens, one per output vertex (features: the vertex's normalized grid coordinates), cross-attends to the contextualized observations, and a decoder MLP maps each attended query to a 3D position. Per-vertex cross-attention preserves locality: each output vertex reads the observations individually rather than a single pooled summary, which makes reconstruction sensitive to touch placement. The network is permutation-invariant and accepts any number of touches. Setting the query grid to (H, W), $(H, 1)$, or (D, H, W) yields the cloth, rope, and soft-body estimators from one implementation.

Belief and learned acquisition: We train M=5 independently initialized estimators. The ensemble mean is the reconstruction and the per-vertex variance is an epistemic uncertainty, high where observations under-constrain the shape. We define an *expected-error-reduction (EER) oracle* that tentatively adds each candidate touch, reconstructs, and measures the resulting true error. It uses ground truth and serves as an upper bound. A deployable agent instead scores candidates from belief features only (ensemble-mean position, ensemble variance, grid position, fraction observed, distance to nearest observation) via a small MLP trained by regression against the oracle on on-policy rollouts with ϵ-greedy exploration (ϵ=0.4). Greedy-only training instead caused a covariate shift: the policy met unfamiliar belief states at test time and fell below random at the largest budgets. Because features are per candidate vertex, the same acquisition transfers across 1D, 2D, and 3D objects. The learned policy amortizes the oracle's per-candidate lookahead into its weights: at deployment each step needs a single ensemble reconstruction plus a cheap scoring pass, fewer network evaluations than the oracle.

IV. Experiments

Setup: We generate configurations in the MuJoCo flex simulator [14]: a 20×20 cloth draped over a randomized obstacle into self-occluded folds, a 40-vertex rope settled into random curves, and a 6×6×6 (216-vertex) volumetric soft body. We generate $\sim 10^3$ configurations per object with a held-out test split, train five-member ensembles (2500 steps, Adam, σ=0.02), and report mean ± std over seeds (three for the estimator, ablation, occlusion, and soft-body results; up to ten for active sensing). For *estimation* we compare inverse-distance weighting (IDW), a Gaussian-process implicit-surface (GPIS) reconstruction, and a global-pool (DeepSets-style [17]) set encoder. For *active sensing* we compare random, a GP uncertainty baseline [16], an ensemble-variance proxy, our learned acquisition, and the EER oracle. Unless noted we report percent error reduction, $100\,(\mathcal{E}_{\text{ref}} - \mathcal{E})/\mathcal{E}_{\text{ref}}$, against the appropriate reference: random touching for active sensing, and the named baseline for estimation.

TABLE I
Estimator error reduction vs. non-learned baselines (%). Budgets low/mid/high/highest are K=10, 15, 20, 30 for cloth and K=4, 6, 8, 12 for rope.

	Cloth		Rope	
Budget	vs. GPIS	vs. IDW	vs. GPIS	vs. IDW
low	+74.8	+66.5	+77.9	+66.6
mid	+73.7	+65.1	+77.4	+58.4
high	+71.8	+65.8	+75.3	+53.9
highest	+66.6	+65.4	+68.7	+45.3

The learned estimator outperforms the non-learned baselines by a wide margin. Table I reports its error reduction over IDW and GPIS at random observation sets, and Fig. 2 shows example reconstructions. Across budgets and both object classes, it cuts error by roughly two-thirds. The margin against GPIS is as large as against IDW: a Gaussian-process surface over-smooths

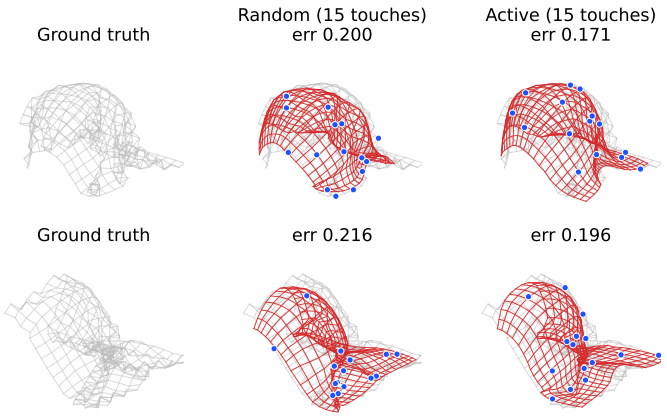

Fig. 2. Qualitative reconstruction of two heavily-crumpled cloth configurations from 15 touches (blue dots). Left: ground truth (grey), middle: reconstruction from random touches, right: from actively-selected touches, with per-vertex error annotated. The estimator recovers global fold structure from sparse contacts, and active selection further reduces error.

TABLE II
ACTIVE SENSING ON CLOTH, % ERROR REDUCTION VS. RANDOM.

K	Variance	GP [16]	Learned	Oracle
10	$+3.3$	-2.7	$+6.2$	$+14.2$
15	-0.1	$+0.4$	$+7.0$	$+20.6$
20	-4.0	$+1.7$	$+5.6$	$+21.4$
30	-11.4	$+5.4$	$+5.8$	$+20.2$

the sharp folds of a crumpled mesh, while cross-attention captures them. When we ablate cross-attention against the global-pool encoder under identical training, cross-attention is consistently better, and the margin grows with budget (from $+4.7\%$ to $+22.4\%$ on cloth as touches accumulate). This says the *locality* of cross-attention, not just the use of a learned set encoder, is what drives reconstruction quality.

The same architecture also handles a volumetric object. With only its query grid changed to $(6, 6, 6)$, it reconstructs the soft body from sparse touch and reduces error over GPIS by $+71$ to $+80\%$, a margin that matches cloth and rope. A single network thus covers 1D, 2D, and 3D objects.

Active sensing gives a real but modest gain, and its size depends on the regime. Table II and Fig. 3 report it on cloth and rope. The learned policy reduces error over random by $+5.6$ to $+7.0\%$ across budgets and beats the GP baseline at sparse budgets. The GP baseline catches up only at the largest budget, where coverage matters more than selection. The ensemble-variance proxy degrades below random as the budget grows, because maximizing variance concentrates touches on persistently uncertain outliers rather than informative locations. The oracle reveals substantial headroom ($+14$ to $+21\%$): the gap between deployable and ideal selection, not the absence of headroom, bounds the gains. The same learned acquisition shows the same structure on the 1D rope and on the 3D soft body ($+17$ to $+18\%$ over random at sparse budgets), supporting the topology-agnostic claim.

The benefit of active sensing grows with occlusion. Stratifying the cloth test set by crumple severity, the learned policy's error reduction rises from $+6.0\%$ on the least-occluded third of configurations to $+11.8\%$ on the most-occluded third, while the oracle headroom rises from $+17.8\%$ to $+22.8\%$ (Fig. 4).

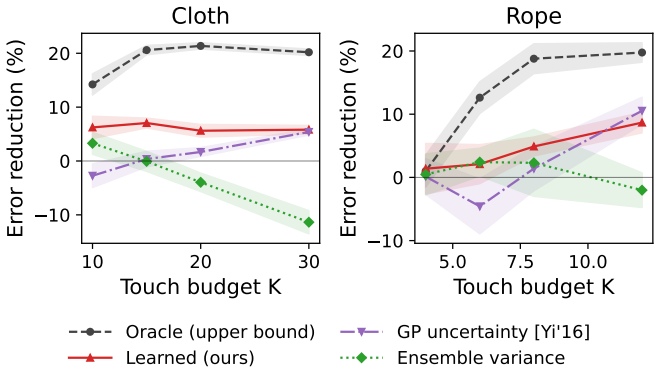

Fig. 3. Active-sensing error reduction over random vs. touch budget, cloth (left) and rope (right), averaged over seeds (shaded: std). The learned policy beats the GP baseline at sparse budgets; the ensemble-variance proxy degrades below random; the oracle bounds the achievable gain.

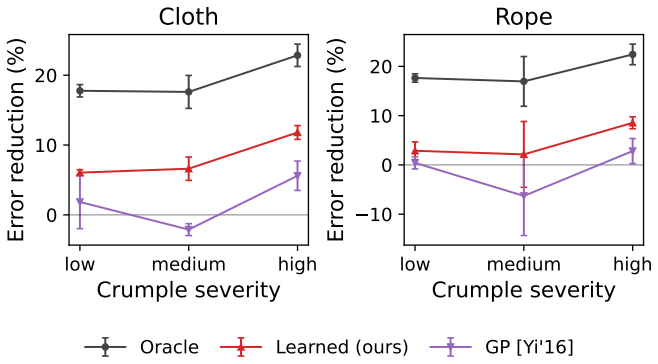

Fig. 4. The active-sensing advantage grows with self-occlusion: stratifying cloth (left) and rope (right) configurations by crumple severity, both the learned policy and the oracle improve most on the most-occluded configurations, the regime where a vision-based prior is least reliable.

A natural question is whether active touch still helps once a camera is available. We find that it largely does not. With a noisy depth camera and a learned shape prior alongside touch, active touch lowers occluded-region error by only about 4 to 6%, and the choice of where to touch barely affects the outcome: random and oracle placement perform almost the same. The camera and the prior already fix most of the hidden state, so a touch has little left to resolve. Without vision the state is underdetermined again, and that is where placement begins to matter.

V. DISCUSSION

These results suggest that the leverage in tactile deformable estimation lies in the estimator, not in where to touch. A locality-aware cross-attention model already recovers most of what sparse random touches allow across 1D, 2D, and 3D objects. Active sensing is a real but bounded add-on whose

benefit concentrates under self-occlusion and vanishes once vision fixes the hidden state. Full-mesh estimates feed directly into deformable planners [13, 7], though we do not evaluate planning here. All experiments are in simulation. The position-only touch model matches a standard arm with descend-until-contact and forward kinematics, so no specialized tactile skin is required for a hardware demo. The main open algorithmic gap is active selection: closing the oracle-deployable divide, especially under realistic probing, via belief-conditioned policies or amortized lookahead over the belief grid.

VI. Conclusion

We presented a topology-agnostic estimator that reconstructs full deformable meshes from sparse touch alone. Learned active sensing helps modestly under occlusion but not when vision is available. Real-robot validation and stronger active policies are the next steps.

References

[1] Cheng Chi and Shuran Song. GarmentNets: Category-level pose estimation for garments via canonical space shape completion. In *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021.

[2] Stanimir Dragiev, Marc Toussaint, and Michael Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 2845–2850, 2011.

[3] Zixuan Huang, Xingyu Lin, and David Held. Mesh-based dynamics with occlusion reasoning for cloth manipulation. In *Proc. Robotics: Science and Systems (RSS)*, 2022.

[4] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[5] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[6] Kangchen Lv, Mingrui Yu, Yifan Pu, Xin Jiang, Gao Huang, and Xiang Li. Learning to estimate 3-D states of deformable linear objects from single-frame occluded point clouds. *arXiv:2210.01433*, 2023.

[7] Xiao Ma, David Hsu, and Wee Sun Lee. Learning latent graph dynamics for visual manipulation of deformable objects. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2022.

[8] Riccardo Mazza, Ciro Natale, and Pietro Falco. Active cross-modal visuo-tactile perception of deformable linear objects. *arXiv:2601.13979*, 2026.

[9] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 441–448, 2001.

[10] Lukáš Rustler, Jens Lundell, Jan Kristof Behrens, Ville Kyrki, and Matej Hoffmann. Active visuo-haptic object shape completion. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[11] Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. Active 3D shape reconstruction from vision and touch. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. NeuralFeels with neural fields: Visuo-tactile perception for in-hand manipulation. *Science Robotics*, 2024.

[13] Tongxuan Tian, Haoyang Li, Bo Ai, Xiaodi Yuan, Zhiao Huang, and Hao Su. Diffusion dynamics models with generative state estimation for cloth manipulation. In *Proc. Conf. Robot Learning (CoRL)*, 2025.

[14] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[16] Zhengkun Yi, Roberto Calandra, Filipe Veiga, Herke van Hoof, Tucker Hermans, Yilei Zhang, and Jan Peters. Active tactile object exploration with Gaussian processes. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pages 4925–4930, 2016.

[17] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.