Agreement Volatility: A Second-Order Metric for Uncertainty Quantification in Surgical Robot Learning

Jordan Thompson, Britton Jordan, Daniel S. Brown, and Alan Kuntz*

Abstract:
Autonomous surgical robots are a promising solution to the increasing demand for surgery amid a shortage of surgeons. Recent work has proposed learning-based approaches for the autonomous manipulation of soft tissue. However, due to variability in aspects such as tissue geometries and stiffnesses, these methods do not always perform well, especially in out-of-distribution settings. To address this challenge, we propose a novel second-order metric for uncertainty quantification, agreement volatility, that enables successful and efficient collaborative handoffs between a human operator and a robot during soft tissue manipulation by allowing the robot to know when to cede control to human operators and when to resume autonomous operation. We validate our approach using the daVinci Research Kit (dVRK) surgical robot to perform risk-aware physical soft tissue manipulation. Our experimental results demonstrate that our proposed agreement volatility metric improves system success rates and leads to a 10% lower reliance on human interventions compared to a variance-only baseline. We further demonstrate the usefulness of our agreement volatility metric as a spatial uncertainty map over geometric point cloud data, enabling uncertainty attribution which provides insight into regions of the input causing uncertainty.

Keywords: Uncertainty Quantification and Attribution, Surgical Robotics

1 Introduction

Autonomous surgical robots have the potential to help solve the growing disparity between the population's need for surgery and the number of available surgeons [1, 2]. However, surgical robot learning and automation is particularly challenging due to the nuanced and risk-sensitive nature of the tasks, the partially observable and deformable nature of the (human tissue) environment, and the scarcity of available data. Because autonomous surgical system failures can be detrimental to patient health, it is crucial that they take into account uncertainty so that these autonomous systems can be risk sensitive and safely cede control to the surgeon before causing any harm. To this end, we aim to develop an interpretable surgical soft tissue manipulation handoff policy between a learned autonomous agent and a surgeon supervisor using uncertainty quantification and attribution. In doing so, our goal is to mitigate the risk of system failures while still offloading the work of soft tissue manipulation to the robot, whenever it is safe to do so.

Our main goal is to enable uncertainty-based human-to-robot and robot-to-human handoffs in the domain of soft tissue manipulation. To this end, we build on recent advances in robot learning for deformable object manipulation. In particular, we extend the recently proposed, state-of-the-art DeformerNet framework [3] which uses large-scale self-supervised training in simulation to

*J. Thompson, B. Jordan, D. S. Brown, and A. Kuntz are with the Robotics Center and the Kahlert School of Computing at the University of Utah, Salt Lake City, UT 84112, USA; (email: {jordan.thompson,alan.kuntz}@utah.edu).

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

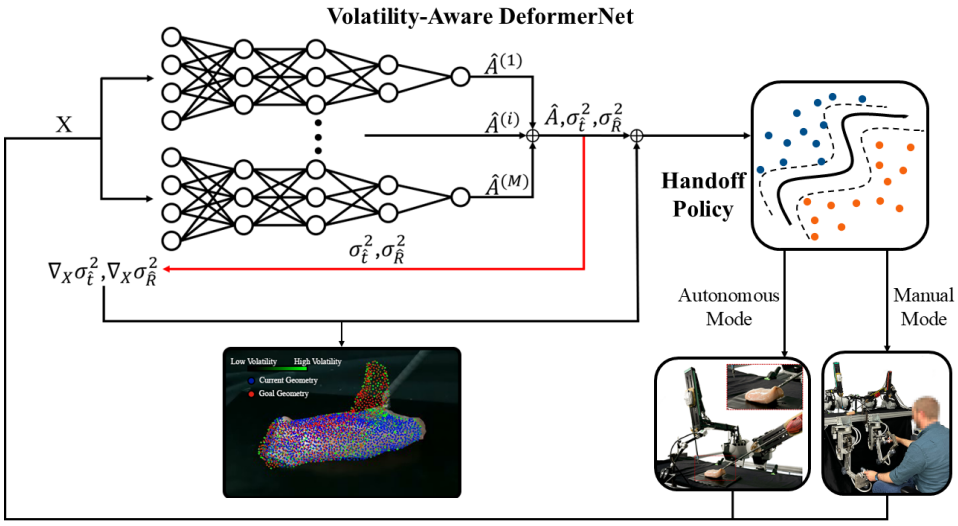

Figure 1: **Volatility-Aware DeformerNet (VAD-Net)** contains an ensemble of DeformerNet [3] architectures trained for manipulating soft tissue. Each model takes as input a current and goal partial-view point cloud of the geometry of the tissue along with the manipulation point and produces a homogeneous transformation matrix $\hat{\mathrm{A}}^{(i)}$ for the end-effector of the robot. From each model's prediction, we produce a predictive distribution containing the final ensemble prediction $\hat{\mathrm{A}}$ along with variance values, $\sigma_{\hat{t}}^2$ and $\sigma_{\hat{R}}^2$, for the positional and rotational components of the set of predictions. These variance values are backpropagated through the ensemble to determine the agreement volatility with respect to the model input, $\nabla_{\mathrm{X}}\sigma_{\hat{t}}^2$ and $\nabla_{\mathrm{X}}\sigma_{\hat{R}}^2$. The variance values and agreement volatilities are used to construct an uncertainty feature set U which are passed through a Support Vector Machine (SVM) handoff policy which determines if the system should request a human to begin teleoperation, or if the system is capable of safely acting autonomously.

learn a closed-loop policy for manipulating deformable objects to desired goal geometries. While DeformerNet has been shown to perform well on in-distribution data, patient-specific anatomical variations, sub-optimal grasping locations, or poorly defined goal geometries can lead to out-of-distribution soft tissue manipulation problems. We show that these out-of-distribution geometries can lead to poor model performance which poses a potential safety risk during surgical manipulation. We propose to address this issue through a novel framework for uncertainty quantification and attribution in surgical robot learning.

Our main contributions are as follows: (1) We propose Volatility-Aware DeformerNet (VAD-Net) the first real-world surgical robotic system capable of uncertainty quantification, attribution, and human-robot collaborative handoffs during soft tissue manipulation. (2) We propose and experimentally validate agreement volatility, a second-order measure of ensemble agreement. We show that agreement volatility provides a more accurate indication of downstream task performance for surgical soft tissue manipulation than using prior ensemble-based uncertainty quantification metrics. (3) We leverage uncertainty attribution to produce spatial uncertainty maps over geometric point cloud data that enables human insight into which region of the geometry has the most influence over the uncertainty.

2 Related Work

Soft Tissue Manipulation: Several data-driven approaches have been proposed to learn soft tissue manipulation [4, 5, 6, 7, 8, 3]. These learning-based approaches are made possible by taking advantage of recent advancements in high-fidelity deformable object simulation [9, 10, 11]. Other work has shown success in using model-independent deformation estimation techniques [12]. Our method takes inspiration from the recently proposed DeformerNet framework [3]. DeformerNet takes a self-supervised learning approach to the problem of soft tissue manipulation. Given the current geometry of the object, the grasping point, and the desired geometry of the object, DeformerNet

has been shown to have state-of-the-art performance for a variety of deformable object geometries and tasks [3]. However, DeformerNet is a black-box policy with no uncertainty quantification and, as we show in our experiments, DeformerNet often fails when given out-of-distribution inputs or inaccurate grasp points. This work, VAD-Net, seeks to remedy these shortcomings by both quantifying the uncertainty during soft tissue manipulation and also developing explainable uncertainty attribution techniques that work with point cloud inputs and deformable object manipulation to enable uncertainty-informed collaborative human-robot handoffs during soft tissue manipulation.

Uncertainty Quantification and Attribution: Uncertainty quantification for deep neural networks has been studied in recent years as a way to aid in the safe design and implementation of deep learning systems [13]. Common approaches for uncertainty estimation include deep ensembles [14] and Monte Carlo (MC) dropout [15, 16, 17]. Ensembles have been shown to improve predictive performance by training multiple models with different random initializations and measuring the variance in their output as a metric of model uncertainty [18, 19, 20] and have been shown to outperform other techniques for uncertainty quantification such as MC dropout [18, 20, 21, 22, 23]. Most techniques for uncertainty quantification rely on first-order metrics such as predictive variance or entropy for quantifying uncertainty [19, 24, 25, 26, 23]. However, while these first-order metrics capture the magnitude of agreement between predictions, they may fail to reflect the stability of model uncertainty given input perturbations. In this work, we propose a novel second-order metric, agreement volatility, which measures the sensitivity of ensemble agreement to input changes. Work has also been done on attributing uncertainty to model inputs. While much of this work has been on gradient-based attribution methods [27, 28, 29, 30], others have proposed the use of input augmentation [31]. Contrary to prior work that primarily focuses on image classification data, we demonstrate that our method attributes uncertainty to geometric point cloud data on a real-world surgical manipulation task. Through these uncertainty maps, we aim to enhance the transparency and trustworthiness of autonomous surgical agents.

Human Interventions for Robot Policies: One of the major obstacles when applying robot learning to surgical domains is the inability to do online learning in the real world. While we can collect demonstrations offline [32, 33, 34, 35, 36, 37, 38, 39] and can train policies in simulation [40, 41, 42, 43], offline imitation learning leads to compounding errors [44] and the sim2real gap is especially difficult to overcome for the types of deformable manipulation that are common in surgery [45]. Enabling an autonomous surgical robotic system to cede control to a surgeon supervisor in high-risk states addresses these challenges by enabling a human to correct the robot during real-world execution. There are two common strategies for deciding when to pass control from the autonomous agent to the human. In the first paradigm, the human decides when to intervene [46, 47, 48]. However, this imposes a large burden on the supervisor. In the second paradigm, which we follow, the robot actively requests human interventions [49, 50, 26, 51] based on some form of uncertainty estimation. However, prior work has focused mainly on simulated tasks with simulated human supervisors and simple control or manipulation tasks. By contrast, we study the efficacy of uncertainty quantification when deployed on a surgical dVRK robot performing real deformable tissue manipulation. In contrast to prior work, we also introduce a novel second-order metric for uncertainty quantification that results in improved performance and more efficient and successful human-robot collaboration and enables uncertainty attribution.

3 Problem Definition

Given a learned surgical robot policy, π_{robot}, for soft tissue manipulation, we aim to develop a risk-sensitive and efficient collaborative handoff framework that enables the robot to cede control to a human supervisor when the probability of failure is high and request control from the human supervisor when the probability of failure is low. Following prior work on deformable tissue manipulation [3], we assume access to a partial-view point cloud of the current geometry of the soft tissue to be manipulated as well as a goal point cloud that defines how the soft tissue should be manipulated. In this work, goal point clouds are generated via manual manipulation of the tissue; however,

these goal point clouds have been shown to be both heuristically definable on a task-specific basis [3] as well as learnable from human demonstrations [52].

We define a meta-policy π_{meta} that operates over uncertainty features to decide whether the robot should continue autonomous operation or request human intervention. This meta-policy can be formalized as a binary classifier $\pi_{\text{meta}} : \mathcal{U} \rightarrow \{0, 1\}$, where $\mathcal{U}$ is any feature space used for measuring uncertainty. We consider true positives (TP) to be when the robot correctly requests an intervention to prevent a task failure; a false positive (FP) occurs when the robot unnecessarily cedes control to the human supervisor. Conversely, false negatives (FN) occur when the robot fails to request an intervention when necessary, and true negatives (TN) occur when the robot successfully manipulates the tissue without requesting an intervention.

Because we are focused on surgical robotic applications, we are most concerned with failures. Thus, we want to minimize failures associated with false negatives. However, false positives may also lead to failure because they unnecessarily distract a human expert and waste time that could be spent addressing other concerns and preventing other failures (e.g., assisting a different patient or robot). To formalize this tradeoff, we define the risk-sensitive objective:

$$\pi^*_{\text{meta}} = \arg\min_{\pi_{\text{meta}}} \mathbb{E}[c_f \cdot FN + c_h \cdot FP], \tag{1}$$

where c_f and c_h are the cost of robot failure and the cost of an unnecessary human handoff respectively. We learn π^*_{meta} as an optimal classification threshold using uncertainty features from a calibration set [53, 54, 55] of real-world autonomous execution data.

4 Methodology

4.1 Volatility-Aware DeformerNet

We propose *Volatility-Aware DeformerNet* (VAD-Net), an extension of DeformerNet [3], a learned closed-loop policy for manipulating deformable objects, that leverages a deep ensemble of feed-forward, self-supervised shape-servoing models for deformable object manipulation with a second-order uncertainty metric, *agreement volatility*, which quantifies the sensitivity of ensemble agreement to input perturbations and enables uncertainty attribution.

VAD-Net takes as input the current and goal geometries of a deformable object, P_c and P_g, represented as partial-view point clouds, along with a designated manipulation point $\boldsymbol{m}$. It outputs a predicted end-effector action $\hat{\mathrm{A}}$ as a homogeneous transformation matrix consisting of $\hat{\mathrm{R}}$, the predicted change in orientation, and $\hat{\boldsymbol{t}}$, the predicted change in position of the end-effector. The input geometries are encoded using PointConv-based feature extractors [56], and the model is trained using data of the form $(\mathrm{P}_c, \mathrm{P}_g, \boldsymbol{m}, \mathrm{A})$, where A is the known transformation applied to the end-effector. We create this training dataset in a self-supervised fashion using Isaac Gym [57] to learn to predict the effects of randomly sampled end-effector manipulation points and manipulation actions on deformable objects given only partial-view point cloud observations. By training an ensemble and computing both predictive variance and agreement volatility, VAD-Net is capable of assessing its own reliability. These uncertainty estimates are used downstream by a learned SVM-based handoff policy (see Section 4.4) to determine when control should be deferred to a human operator. We train VAD-Net consisting of five DeformerNet policies. We use the source code and training and testing datasets provided by Thach et al. [3], consisting of 11,566 training and 1,285 test examples of deformable object manipulations. Following prior work [14, 20, 18], each ensemble component model was trained using the same training dataset with different random weight initializations. See Appendix A in the supplement for more details.

4.2 Uncertainty Quantification

We seek to enable more risk-aware and efficient autonomous handoffs in deformable soft tissue manipulation. Following prior work on ensemble uncertainty [14, 18, 19, 20], we can simply compute the variance across ensemble predictions as a first-order measure of ensemble agreement; however,

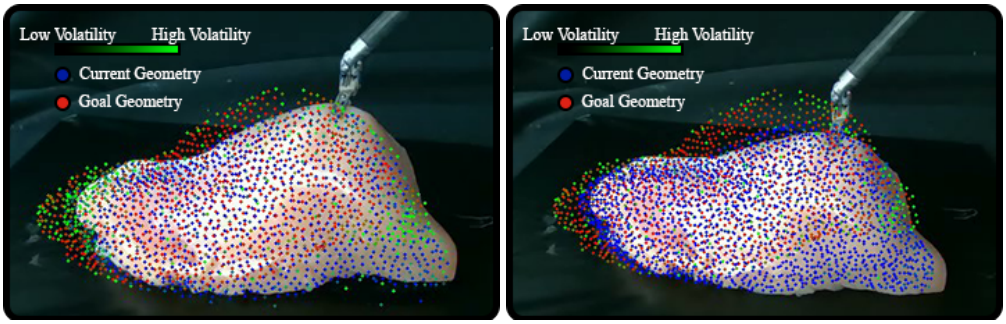

Figure 2: An example of uncertainty attribution on the dVRK experimental setup. The blue points represent the current point cloud of the chicken tissue geometry while the red points represent the goal point cloud geometry. The green intensity indicates regions of higher agreement volatility scores, highlighting their contribution to the ensemble's uncertainty. (left) Spatial uncertainty map using positional agreement volatility. (Right) Spatial Uncertainty map using rotational agreement volatility.

we also propose and derive a second-order sensitivity metric called *agreement volatility* that enables a more detailed and comprehensive measure of uncertainty and naturally lends itself to interpretability via uncertainty attribution. Agreement volatility measures the local sensitivity of ensemble variance with respect to the model's input. High agreement volatility indicates that small perturbations to the input can lead to large fluctuations in predictive confidence, signaling fragile or unreliable regions of the input space. This is particularly important in the context of deformable object manipulation as minor geometric variations caused by sensor noise, occlusions, or slight deformations can pose a risk for reliable task execution.

Our ensemble of M models $\{f^{(1)}, \ldots, f^{(M)}\}$ produces a set of 4×4 homogeneous transformation matrices consisting of a predicted rotation $\hat{\mathrm{R}}^{(i)}$ and a predicted translation $\hat{\boldsymbol{t}}^{(i)}$,

$$f^{(i)}(\mathrm{X}) = \begin{bmatrix} \hat{\mathrm{R}}^{(i)}(\mathrm{X}) & \hat{\boldsymbol{t}}^{(i)}(\mathrm{X}) \\ \mathbf{0} & 1 \end{bmatrix}$$

where X is the model input containing the current and goal point clouds (P_c and P_g) of the tissue geometry along with a manipulation point $\boldsymbol{m}$. We measure uncertainty for both the positional and rotational components.

Positional Volatility: We compute the ensemble's positional prediction as the arithmetic mean of each model's prediction. $\hat{\boldsymbol{t}}(\mathrm{X}) = \frac{1}{M}\sum_{i=1}^{M} \hat{\boldsymbol{t}}^{(i)}(\mathrm{X})$. The ensemble agreement with respect to the positional prediction is measured as the mean squared deviation of $\hat{\boldsymbol{t}}^{(i)}$ from $\hat{\boldsymbol{t}}$:

$$\sigma^2_{\hat{\boldsymbol{t}}}(\mathrm{X}) = \frac{1}{M}\sum_{i=1}^{M} \|\hat{\boldsymbol{t}}^{(i)}(\mathrm{X}) - \hat{\boldsymbol{t}}(\mathrm{X})\|_2^2 \,. \tag{2}$$

We define the agreement volatility as the gradient of the ensemble's variance with respect to the model input $X = (\mathrm{P}_c, \mathrm{P}_g, \boldsymbol{m})$.

$$\nabla_{\mathrm{X}}\sigma^2_{\hat{\boldsymbol{t}}}(\mathrm{X}) = \frac{2}{M}\sum_{i=1}^{M} (\hat{\boldsymbol{t}}^{(i)}(\mathrm{X}) - \hat{\boldsymbol{t}}(\mathrm{X}))^T (\nabla_{\mathrm{X}}\hat{\boldsymbol{t}}^{(i)}(\mathrm{X}) - \nabla_{\mathrm{X}}\hat{\boldsymbol{t}}(\mathrm{X})) \tag{3}$$

This gradient characterizes the local stability of the ensemble's agreement in response to input changes. We then define an agreement volatility score for each point in the input point clouds by taking the ℓ_2-norm of the gradient vector at each input point, yielding a distribution of agreement volatility scores:

$$V_{\hat{\boldsymbol{t}}}(\boldsymbol{p}) = \|\nabla_{\mathrm{X}}\sigma^2_{\hat{\boldsymbol{t}}}(\mathrm{X})|_{\boldsymbol{p}}\|_2, \quad \forall \boldsymbol{p} \in \mathrm{X} \,. \tag{4}$$

Rotational Volatility: Computing analogous metrics for the rotational component of the ensemble's predictions is less straightforward due to the specific structure of rotation matrices. We first compute the arithmetic mean S of the set of rotation matrices produced by the ensemble,

$\mathrm{S}(\mathrm{X}) = \frac{1}{M}\sum_{i=1}^{M} \hat{\mathrm{R}}^{(i)}(\mathrm{X})$. To obtain a valid rotation matrix, we take the singular value decomposition, $\mathrm{S}(\mathrm{X}) = \mathrm{UDV}'$, and multiply U and V′ to obtain a rotation matrix that minimizes the Euclidean norm to S and is therefore the rotation matrix that minimizes the average geodesic distance to the set of rotations produced by the ensemble [58]: $\hat{\mathrm{R}}(\mathrm{X}) = \mathrm{UV}'$. The ensemble agreement with respect to the rotational prediction is then measured as the mean geodesic distance between $\hat{\mathrm{R}}^{(i)}$ and $\hat{\mathrm{R}}$:

$$\sigma^2_{\hat{\mathrm{R}}}(\mathrm{X}) = \frac{1}{M}\sum_{i=1}^{M} \arccos\Big(\frac{\mathrm{Tr}(\hat{\mathrm{R}}^{(i)}(\mathrm{X})\hat{\mathrm{R}}(\mathrm{X})^T) - 1}{2}\Big) \tag{5}$$

Similarly to the positional case, we compute the gradient of the rotational variance with respect to the input point clouds to obtain the rotational agreement volatility:

$$\nabla_{\mathrm{X}}\sigma^2_{\hat{\mathrm{R}}}(\mathrm{X}) = \frac{-1}{2M}\sum_{i=1}^{M} \frac{\nabla_{\mathrm{X}}\mathrm{Tr}(\hat{\mathrm{R}}^{(i)}(\mathrm{X})\hat{\mathrm{R}}(\mathrm{X})^T)}{\sqrt{1 - (\frac{\mathrm{Tr}(\hat{\mathrm{R}}^{(i)}(\mathrm{X})\hat{\mathrm{R}}(\mathrm{X})^T)-1}{2})^2}} \tag{6}$$

Using the same method as for the positional component, we compute the per-point agreement volatility score of the gradient vector at each input point:

$$V_{\hat{\mathrm{R}}} = \|\nabla\sigma^2_{\hat{\mathrm{R}}}(\mathrm{X})|_{\boldsymbol{p}}\|_2, \;\; \forall \boldsymbol{p} \in \mathrm{X}\,. \tag{7}$$

which produces a distribution of agreement volatility scores across the input point clouds.

4.3 Uncertainty Attribution

The variance and agreement volatility metrics described above not only serve as indicators of predictive uncertainty, but also enable attribution of that uncertainty to specific regions of the input. Since both the positional and rotational agreement volatility scores are computed per input point, the ensemble effectively produces an uncertainty map over the current and goal point clouds. Points with a high agreement volatility score indicate regions with high influence over the ensemble's predicted action and confidence. Fig. 2 shows an example of this attribution performed on the dVRK with ex vivo chicken muscle tissue.

4.4 Collaborative Handoff Policy

To learn the meta-policy π^*_{meta} introduced in Sec. 3, we train a Support Vector Machine (SVM) as a binary classifier over the uncertainty feature space $\mathcal{U}$, which includes both ensemble variance and agreement volatility. We choose to use an SVM in this case due to the limited availability of data and their low computational cost. This classifier learns a decision boundary that approximates π^*_{meta}, separating successful and failed trials using features computed from autonomous execution data. We use this decision boundary as a handoff policy that determines when control should be passed between the autonomous agent and the human teleoperator.

Given a dataset of completed trials D, each labeled with task success based on a Chamfer distance [59] threshold, we extract uncertainty features and train the SVM to classify each trial outcome. Chamfer distance acts as a proxy for how accurately the tissue was manipulated to the goal. During deployment, the trained classifier serves as the real-time decision mechanism: if the uncertainty features at the current timestep indicate likely failure, control is given to the human operator. Conversely, if the prediction indicates success, the robot either retains or reclaims control. As recommended by Hoque et al. [26], we implement a hysteretic switching policy that requires a higher confidence for handing control to the robot than for handing control to the human to prevent thrashing between robot and human control. This band introduces stability by requiring the system to accumulate sufficient confidence before autonomous operation, helping to ensure smoother transitions between human and robot control. The output of the SVM handoff policy is a probability that the given task will succeed. If the human is currently teleoperating the robot, this probability must increase above 60% before control will be handed back to the robot. This prevents unnecessary thrashing of control between the human and the robot while also acting as a conservative handoff policy. This framework enables a collaborative policy that balances robustness (false negatives) with efficiency (false positives), aligning with the risk-sensitive objective defined in Eq. (1).

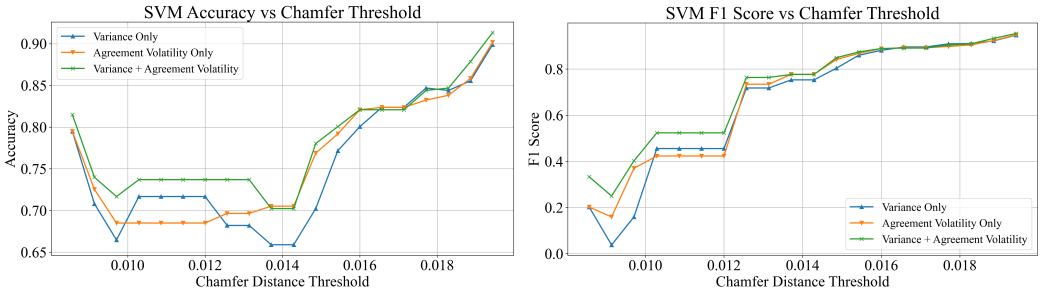

Figure 3: SVM accuracy and F_1-score results across varying Chamfer distance success thresholds. We show results using only variance, only agreement volatility, and both variance and agreement volatility. We find that the introduction of agreement volatility in general improves both the accuracy and the F_1-score of the classifier.

5 Physical Experimental Setup

The deep ensemble is implemented using a zero-shot sim2real framework. Trained entirely on a simulated box-shaped deformable object, the system is tasked with manipulating ex vivo chicken tissue of varying geometries. To generate goal shapes for evaluation purposes only, using ex vivo tissue we teleoperated the robot to manually manipulate the tissue to a desired goal geometry. We then reset the system and task the robot with manipulating the tissue to the same desired geometry (with no knowledge of how the shape was generated). We use an Intel Realsense D405 camera for tracking point cloud representations of the tissue geometry both in goal generation for evaluation and during method execution. We track the ensemble variance and agreement volatility as the robot manipulates the tissue toward the goal geometry and measure task success as whether the Chamfer distance between the final tissue geometry and the goal geometry is below a predefined threshold. We define a termination criterion for the method as when $\|\hat{\boldsymbol{p}}\| < 0.001$ as this is when there are no more substantial deformations that occur in the tissue. We implement this system on the daVinci Research Kit (dVRK) surgical robot [60] using the patient-side manipulators for tissue manipulation and the surgeon-side console for teleoperation.

6 Physical soft tissue Manipulation Results

6.1 Performance Validation

Prior work shows DeformerNet to be capable of 100% success rates on in-distribution cases [3]. To validate that VAD-Net achieves similar performance, we performed 20 in-distribution trials and achieved a 100% success rate across these cases. However, to highlight the need for a collaborative handoff policy, we also performed 15 out-of-distribution trials across 3 cases: Bad Manipulation Point, OOD Geometry, and Nonlocal Control. VAD-Net (without uncertainty quantification) failed to complete the task in all but 1 of these trials, highlighting the need for human intervention in OOD cases.

6.2 Uncertainty Quantification

We collected 40 trials that contain both in-distribution and out-of-distribution cases of DeformerNet that act on ex vivo chicken muscle tissue and measured the predictive variances and agreement volatilities across the trials using the methodology in Sec. 4.2.

Handoff Policy Calibration Using these 40 trials as a calibration dataset, we trained an SVM across various Chamfer distance success thresholds. We defined 3 distinct uncertainty feature sets containing variance values from Eqs. (2) and (5) as well as the median and inter-quartile range from the agreement volatility distributions found from Eqs. (4) and (7). Fig. 3 shows the accuracy and F_1-score for SVMs trained on each feature set. We find that the introduction of agreement volatility improves both the accuracy and the F_1-score of the SVM from the variance only baseline, illustrating how our novel, second-order metric serves to improve the reliability of a learned handoff policy.

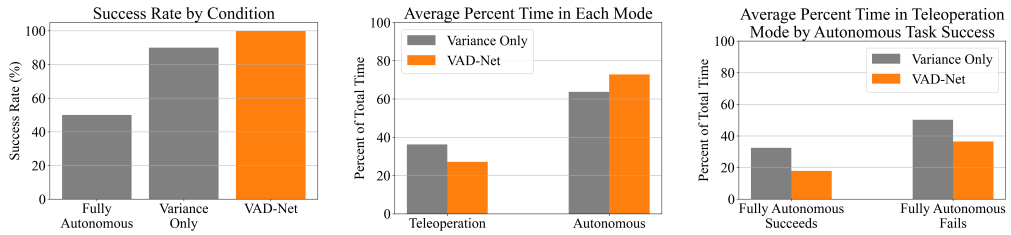

Figure 4: Results from deploying our handoff policy on the dVRK surgical robot for a soft tissue manipulation task. For each trial, we gathered results across three conditions: Fully Autonomous, Variance Only, and VAD-Net, where the latter two utilized the corresponding SVM to trigger handoffs between a human operator and the autonomous agent. (Left) Task success rates across each condition. (Middle) Percent of total task time spent in each operating mode. (Right) Percent of total time spent teleoperating split across cases where the Fully Autonomous system succeeded and failed. The left bars show cases where interventions were requested but unnecessary while the right bars show that in cases where interventions were necessary (fully autonomous robot fails), VAD-Net requires less intervention than the baseline.

Handoff Policy Evaluation We compared the performance of an SVM trained on raw variance values to one augmented with agreement volatility metrics across a set of 30 test trials. Each trial had one of three conditions: Fully Autonomous, Variance Only, and VAD-Net, where the latter two utilized the corresponding SVM to trigger handoffs between a human operator and the autonomous agent. A trial was considered successful if the Chamfer distance between the geometry of the tissue and the goal geometry fell below 1 cm.

Fig. 4 summarizes the results. As expected, the Fully Autonomous baseline condition had the lowest success rate, reinforcing the need for a collaborative handoff policy. VAD-Net achieved a 100% success rate, an improvement over the Variance Only baseline. VAD-Net also spends approximately 10% less time on average in teleoperation mode than the Variance Only baseline, indicating that VAD-Net more efficiently utilizes human interventions. In cases where the Fully Autonomous system failed, VAD-Net required 15% less time teleoperating than the baseline while still achieving a 100% success rate. While the baseline spends more time teleoperating, in failure cases, the baseline never requested an intervention. This highlights VAD-Net's ability to efficiently utilize human interventions to facilitate task success while maximizing the time spent autonomously operating.

Uncertainty Attribution In addition to improved success rates and reduced teleoperation time, our method produces real-time interpretable spatial attributions over the tissue geometry (Fig. 2) at a rate of 20 Hz. These highlight regions of the geometry that contribute the most to model uncertainty, allowing the system to not only decide when to hand off control, but also provide some level of insight into why the system is uncertain. Fig. 2 shows an example of this attribution using both positional and rotational agreement volatility. See Appendix C for more detailed examples of uncertainty attribution being applied during real-time execution.

7 Conclusion

In this work, we presented VAD-Net, a volatility-aware extension of DeformerNet that enables real-time uncertainty estimation and interpretable collaborative control for surgical soft tissue manipulation. By combining ensemble variance with a novel second-order metric, agreement volatility, our system can preemptively identify unstable predictions and trigger handoffs between a human and the autonomous system. Through experiments on physical soft tissue using the dVRK surgical robot, we demonstrated that incorporating agreement volatility improves the task success rate of the system and reduces reliance on human interventions compared to a baseline handoff policy using only ensemble variance. We further introduced a spatial attribution method that highlights the geometric regions contributing most to predictive uncertainty, allowing for more transparent and explainable robot behavior. These results highlight the potential of uncertainty-aware learning to improve both system safety and trust in high-stakes surgical robotic manipulation tasks.

8 Limitations

While our results demonstrate the effectiveness of VAD-Net for collaborative surgical manipulation, several limitations remain for future work. Our work focuses on a complex, real-world soft tissue manipulation task with real human interventions. Due to the time required for this kind of complex evaluation, we only evaluated our method on a single ex vivo tissue task. In future work, we plan to extend this evaluation to more diverse surgical tasks and settings. Additionally, our handoff policy relies on fixed thresholds for Chamfer distance and hysteresis, which may limit adaptability across different task domains or robot platforms. A more principled approach to threshold selection, such as learning thresholds from task-level outcomes, could improve generalization.

Acknowledgments

Research reported in this publication was supported by the Advanced Research Projects Agency for Health (ARPA-H) under Award Number D24AC00415-00. The ARPA-H award provided 100% of the total costs with an award total of up to $11,935,038. The content is solely the responsibility of the authors and does not necessarily represent the official views of ARPA-H.

References

[1] X. Zhang, D. Lin, H. Pforsich, and V. W. Lin. Physician workforce in the united states of america: forecasting nationwide shortages. *Human resources for health*, 18(1):1–9, 2020.

[2] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri. Autonomy in surgical robotics. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:651–679, 2021.

[3] B. Thach, B. Y. Cho, T. Hermans, and A. Kuntz. Deformernet: Learning bimanual manipulation of 3d deformable objects. *arXiv preprint arXiv:2305.04449*, 2023.

[4] A. Attanasio, B. Scaglioni, M. Leonetti, A. F. Frangi, W. Cross, C. S. Biyani, and P. Valdastri. Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study. *IEEE Robotics and Automation Letters*, 5(4):6528–6535, 2020.

[5] A. Pore, E. Tagliabue, M. Piccinelli, D. Dall'Alba, A. Casals, and P. Fiorini. Learning from demonstrations for autonomous soft-tissue retraction. In *2021 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2021.

[6] C. Shin, P. W. Ferguson, S. A. Pedram, J. Ma, E. P. Dutson, and J. Rosen. Autonomous tissue manipulation via surgical robot using learning based model predictive control. In *2019 International conference on robotics and automation (ICRA)*, pages 3875–3881. IEEE, 2019.

[7] S. A. Pedram, P. W. Ferguson, C. Shin, A. Mehta, E. P. Dutson, F. Alambeigi, and J. Rosen. Toward synergic learning for autonomous manipulation of deformable tissues via surgical robots: An approximate q-learning approach. In *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pages 878–884. IEEE, 2020.

[8] M. Retana, K. Nalamwar, D. T. Conyers, S. F. Atashzar, and F. Alambeigi. Autonomous data-driven manipulation of an unknown deformable tissue within constrained environments: A pilot study. In *2022 International Symposium on Medical Robotics (ISMR)*, pages 1–7. IEEE, 2022.

[9] K. Erleben. Non-smooth newton methods for deformable multi-body dynamics. 2019.

[10] I. Huang, Y. Narang, C. Eppner, B. Sundaralingam, M. Macklin, T. Hermans, and D. Fox. Defgraspsim: Simulation-based grasping of 3d deformable objects. *arXiv preprint arXiv:2107.05778*, 2021.

[11] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox. Gpu-accelerated robotic simulation for distributed reinforcement learning. In *Conference on Robot Learning*, pages 270–282. PMLR, 2018.

[12] F. Alambeigi, Z. Wang, Y.-h. Liu, R. H. Taylor, and M. Armand. Toward semi-autonomous cryoablation of kidney tumors via model-independent deformable tissue manipulation technique. *Annals of biomedical engineering*, 46:1650–1662, 2018.

[13] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

[14] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, and D. Batra. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*, 2015.

[15] D. Milanés-Hermosilla, R. Trujillo Codorniú, R. López-Baracaldo, R. Sagaró-Zamora, D. Delisle-Rodriguez, J. J. Villarejo-Mayor, and J. R. Núñez-Álvarez. Monte carlo dropout for uncertainty estimation and motor imagery classification. *Sensors*, 21(21):7241, 2021.

[16] R. Camarasa, D. Bos, J. Hendrikse, P. Nederkoorn, E. Kooi, A. Van Der Lugt, and M. De Bruijne. Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, pages 32–41. Springer, 2020.

[17] Y. Mae, W. Kumagai, and T. Kanamori. Uncertainty propagation for dropout-based bayesian neural networks. *Neural Networks*, 144:394–406, 2021.

[18] S. Fort, H. Hu, and B. Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[19] T. Pearce, A. Brintrup, M. Zaki, and A. Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR, 2018.

[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[21] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[22] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

[23] D. Shin, A. D. Dragan, and D. S. Brown. Benchmarks and algorithms for offline preference-based reward learning. *Transactions on Machine Learning Research*, 2023.

[24] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[25] J. Mena, O. Pujol, and J. Vitria. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35, 2021.

[26] R. Hoque, A. Balakrishna, E. Novoseller, A. Wilcox, D. S. Brown, and K. Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. *arXiv preprint arXiv:2109.08273*, 2021.

[27] H. Wang, D. Joshi, S. Wang, and Q. Ji. Gradient-based uncertainty attribution for explainable bayesian deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12044–12053, 2023.

[28] H. Wang, S. Wang, and Q. Ji. Semantic attribution for explainable uncertainty quantification. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pages 101–112. Springer, 2023.

[29] I. Perez, P. Skalski, A. Barns-Graham, J. Wong, and D. Sutton. Attribution of predictive uncertainties in classification models. In *Uncertainty in Artificial Intelligence*, pages 1582–1591. PMLR, 2022.

[30] B. Kantz, S. Steger, C. Staudinger, C. Feilmayr, J. Wachlmayr, A. Haberl, S. Schuster, and F. Pernkopf. Input uncertainty attribution by uncertainty propagation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[31] S. Jha, S. Raj, S. Fernandes, S. K. Jha, S. Jha, B. Jalaian, G. Verma, and A. Swami. Attribution-based confidence metric for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1), 1991.

[33] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 7 2018.

[34] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*, 2018.

[35] S. Arora and P. Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *arXiv preprint arXiv:1806.06877*, 2018.

[36] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2), 2013.

[37] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann. Probabilistic movement primitives. In *Advances in neural information processing systems*, pages 2616–2624, 2013.

[38] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR, 2022.

[39] W. Zhang, H. Xu, H. Niu, P. Cheng, M. Li, H. Zhang, G. Zhou, and X. Zhan. Discriminator-guided model-based offline imitation learning. In *Conference on Robot Learning*, pages 1266–1276. PMLR, 2023.

[40] W. Zhao, J. P. Queralta, and T. Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, pages 737–744. IEEE, 2020.

[41] A. Pashevich, R. Strudel, I. Kalevatykh, I. Laptev, and C. Schmid. Learning to augment synthetic images for sim2real policy transfer. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2651–2657. IEEE, 2019.

[42] M. Kaspar, J. D. M. Osorio, and J. Bock. Sim2real transfer for reinforcement learning without dynamics randomization. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4383–4388. IEEE, 2020.

[43] S. Sharma, E. Novoseller, V. Viswanath, Z. Javed, R. Parikh, R. Hoque, A. Balakrishna, D. S. Brown, and K. Goldberg. Learning switching criteria for sim2real transfer of robotic fabric manipulation policies. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 1116–1123. IEEE, 2022.

[44] S. Ross, G. J. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[45] M. Haiderbhai, R. Gondokaryono, T. Looi, J. M. Drake, and L. A. Kahrs. Robust sim2real transfer with the da vinci research kit: A study on camera, lighting, and physics domain randomization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3429–3435. IEEE, 2022.

[46] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2019.

[47] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. In *Proc. Robotics: Science and Systems (RSS)*, 2020.

[48] A. Mandlekar, D. Xu, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese. Human-in-the-loop imitation learning using remote teleoperation, 2020.

[49] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer. EnsembleDAgger: A Bayesian Approach to Safe Imitation Learning. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[50] J. Zhang and K. Cho. Query-efficient imitation learning for end-to-end autonomous driving. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2017.

[51] G. Datta, R. Hoque, A. Gu, E. Solowjow, and K. Goldberg. Iifl: Implicit interactive fleet learning from heterogeneous human supervisors. In *Conference on Robot Learning*, pages 2340–2356. PMLR, 2023.

[52] B. Thach, T. Watts, S.-H. Ho, T. Hermans, and A. Kuntz. Defgoalnet: Contextual goal learning from demonstrations for deformable object manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3145–3152. IEEE, 2024.

[53] P. A. Flach. Roc analysis. In *Encyclopedia of machine learning and data mining*, pages 1–8. Springer, 2016.

[54] O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems*, 27, 2014.

[55] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[56] W. Wu, Z. Qi, and L. Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019.

[57] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[58] S. Sarabandi, A. Shabani, J. M. Porta, and F. Thomas. On closed-form formulas for the 3-d nearest rotation matrix problem. *IEEE Transactions on Robotics*, 36(4):1333–1339, 2020.

[59] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, and D. Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 29088–29100, 2021.

[60] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio. An open-source research kit for the da vinci surgical system. In *IEEE Intl. Conf. on Robotics and Auto. (ICRA)*, pages 6434–6439, Hong Kong, China, 2014.

[61] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

Appendix A DeformerNet Details

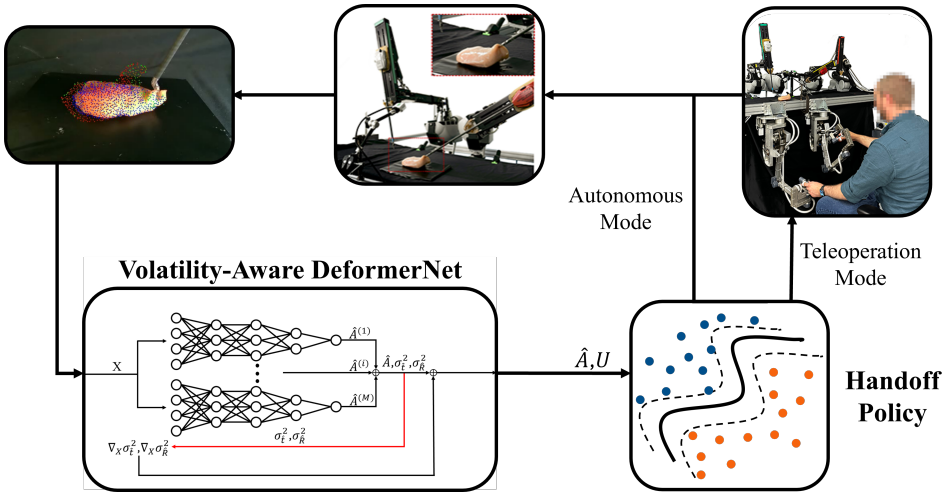

Figure 5: **Example VAD-Net Execution**

Our VAD-Net method (an example execution of which is shown in Fig. 5) is an ensemble consisting of five independently initialized instances of the DeformerNet architecture. The original DeformerNet architecture is composed of two PointConv-based feature extractors that independently process the current and goal geometries, P_{cm} and P_g, both represented as partial-view point clouds of the deformable object. Each point cloud initially has the shape 1024×3, corresponding to 1024 3D points. To encode task-relevant context, the current geometry P_c is augmented with the manipulation point $\boldsymbol{m}$ to produce $\mathrm{P}_{cm} \in \mathbb{R}^{1024\times 4}$, where the first three channels encode the spatial coordinates of each point, and the fourth channel is a binary indicator marking the 50 points nearest to $\boldsymbol{m}$. This augmented input enables the model to focus on regions relevant to the planned manipulation.

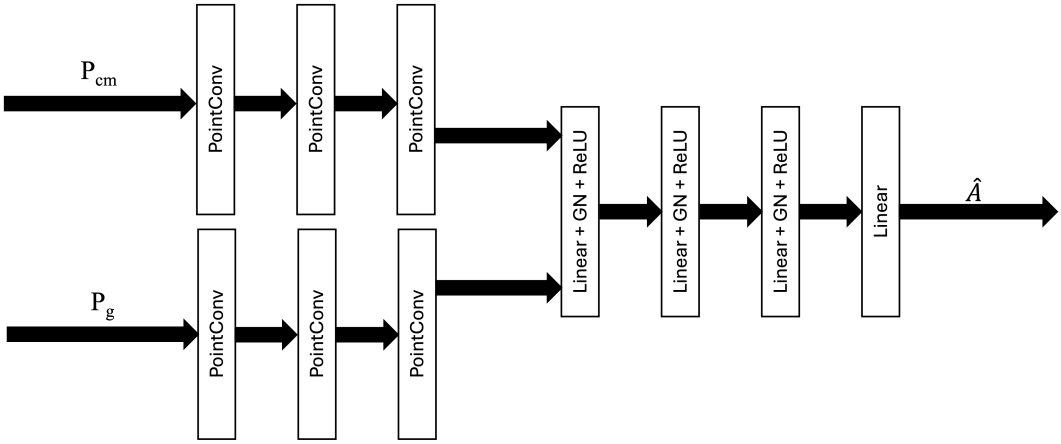

Figure 6: **DeformerNet Architecture**

In our VAD-Net ensemble, each of the five independently initialized instances of the DeformerNet architecture outputs a predicted 4×4 transformation matrix $\hat{\mathrm{A}}$ composed of a rotation matrix $\hat{\mathrm{R}} \in \mathrm{SO}(3)$, and a translation vector $\hat{\boldsymbol{t}} \in \mathbb{R}^3$.

Fig. 6 shows a visualization of the full DeformerNet architecture. The feature extractors contain three sequential PointConv layers with increasing feature dimensions of 64, 128, and 256. The encodings of P_{cm} and P_g are then concatenated into a vector of length 512. This vector is then passed through a sequence of four feedforward fully connected layers separated by a GroupNorm

layer and a ReLU activation function to produce a 9-dimensional output vector containing $\hat{\boldsymbol{t}}$ and a 6D vector which gets mapped to a rotation matrix $\hat{\mathrm{R}}$ using the method described in Zhou et al. [61].

Training data is collected entirely in Isaac Gym [57] on a deformable box object. Given a current geometry of the object P_c and a random manipulation point $\boldsymbol{m}$, a random action A is applied to the robot's end-effector. The final geometry of the object is recorded as P_g. The model is then trained on tuples of the form $(\mathrm{P}_c, \mathrm{P}_g, \boldsymbol{m}, \mathrm{A})$ where A contains the ground-truth translation $\boldsymbol{t}$ and rotation R. The model loss is a linear combination of the mean squared error between the $\hat{\boldsymbol{t}}$ and $\boldsymbol{t}$ and the geodesic distance between $\hat{\mathrm{R}}$ and R.

We train the model using an Adam optimizer with a learning rate of 0.001 over 200 epochs using a dataset containing 11,566 training and 1,285 test examples. After 100 epochs, the learning rate is reduced to 0.0001.

Appendix B Additional Handoff Trial Results

To further analyze the performance of our uncertainty-aware handoff framework, we include additional experimental results that provide deeper insight into the relationship between predictive uncertainty and task success, as well as the efficacy of the handoff policies.

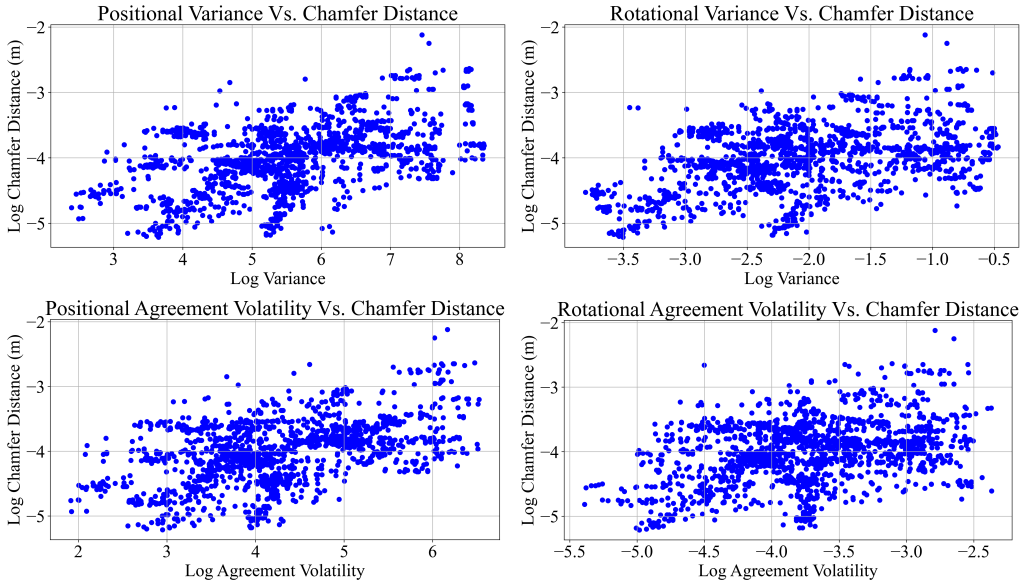

Figure 7: **Uncertainty Metric Correlations**

Correlation Between Uncertainty Metrics and Task Error Fig. 7 shows log plots of the correlation between each of our uncertainty metrics (ensemble variance and agreement volatility) and the Chamfer distance between the current and goal tissue geometries after taking the predicted action. Across both positional and rotational components, we observe a positive linear relationship. This supports our hypothesis that agreement volatility captures higher-order information about predictive stability and is indicative of downstream task success. These results reinforce the decision to use agreement volatility as an uncertainty feature in the learned handoff policy.

Task Efficiency Across Conditions Fig. 8 compares the relative task efficiency between the Variance Only and VAD-Net policies. For each trial in our 30-trial test set, we record which condition (Variance Only or VAD-Net) completed the task in less total time. We report the percentage of trials in which each method was faster. VAD-Net completed the task faster in approximately 80% of cases (with an average reduction in time of 32.07%), demonstrating its ability to operate more efficiently than the Variance Only baseline. This result highlights the practical advantage of incorporating agreement volatility into the handoff-policy, enabling the system to operate more efficiently while still maintaining high success rates.

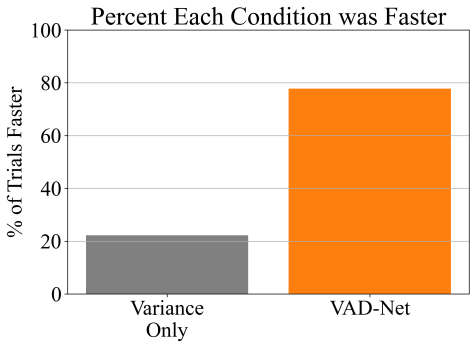

Figure 8: **Relative Condition Speed**

Table 1 presents a summary of key performance metrics across the three experimental conditions. As expected, the Fully Autonomous baseline

Table 1: Comparison of key performance metrics across control conditions.

Metric	Fully Autonomous	Variance Only	VAD-Net
Success Rate (%)	50	90	100
Avg. Teleoperation Time / Handoff (s)	–	12.197	4.571
Avg. Autonomous Time / Handoff (s)	–	16.249	16.566
% Teleoperation	–	36.292	27.170
% Autonomous	–	63.708	72.83
Avg. # of Handoffs	–	2	2.1

achieves the lowest task success rate, reinforcing the need for uncertainty-aware handoffs. The Variance Only policy improves the success rate to 90%, but still requires substantial human intervention with an average of 12.2 seconds spent in teleoperation mode and 2.0 handoffs per trial. In contrast, VAD-Net achieves a 100% success rate while reducing average teleoperation time to just 4.6 seconds, indicating greater trust in the policy's predictions. Despite a similar number of handoffs, the more efficient decision-making enabled by agreement volatility reduces overall reliance on human intervention. We note that the raw number of seconds spent in teleoperation/autonomous mode is not as interpretable as the percentages as there can be a wide range of variance between the length of each trial. Thus, the percentages normalize these values to be independent of the length of the trial. These results demonstrate that VAD-Net not only enhances reliability in challenging manipulation tasks, but also enables more autonomous operation without compromising safety.

Monte Carlo Dropout As an additional baseline for comparison, we also implemented a Monte Carlo Dropout (MC Dropout) version of VAD-Net. However, in testing this version of VAD-Net (without uncertainty quantification) we found that in 10 *in-distribution* trials, this model failed to manipulate the tissue in 9 out of 10 trials. This illustrates that the introduction of dropout to the network severely hinders model performance. Thus, this MC Dropout version was not used throughout our experiments.

Appendix C Uncertainty Attribution Examples

In this section, we present additional insight into VAD-Net's uncertainty attribution mechanism during real-world execution on the dVRK. VAD-Net generates spatial uncertainty maps that highlight regions in the model input that have the greatest influence on the model's uncertainty, enabling both interpretability and collaborative handoffs.

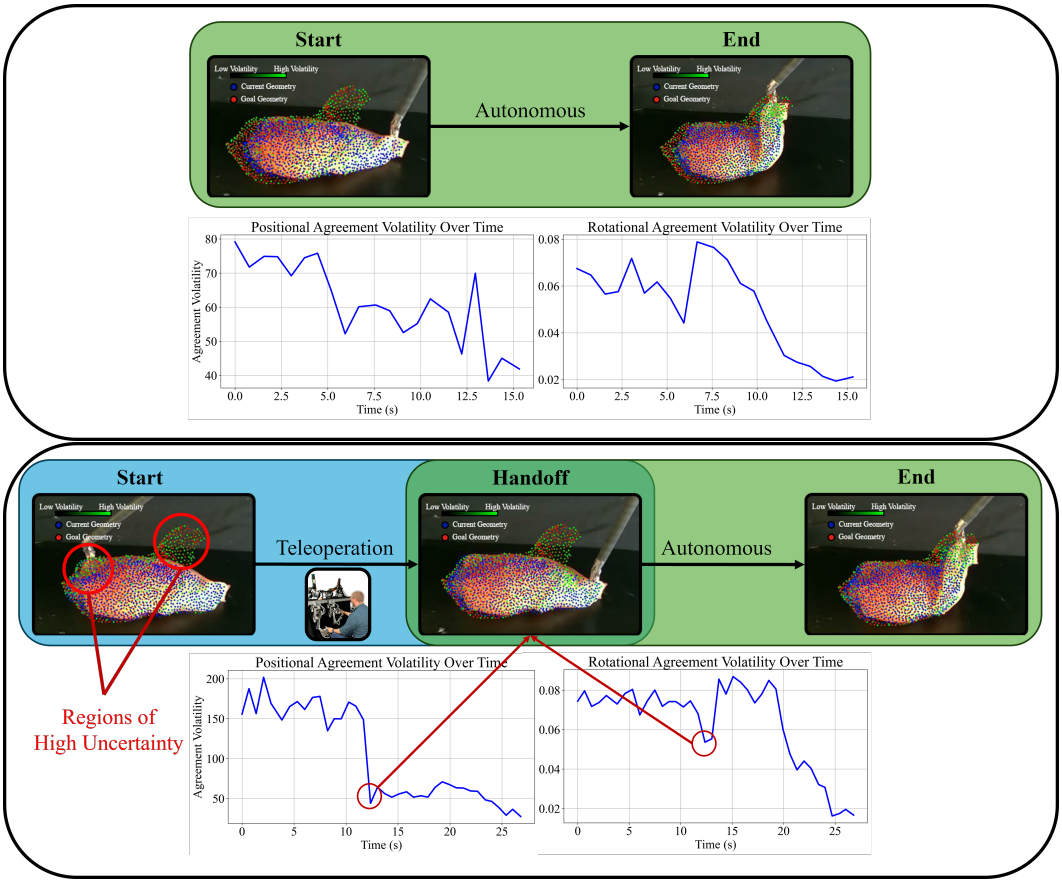

Figure 9: **VAD-Net Trial Examples**

Case 1: Fully Autonomous Execution In the first example (top row of Fig. 9), VAD-Net successfully operates fully autonomously without the need for human intervention. Success without the need for human intervention is currently atypical in out-of-distribution scenarios, as mentioned above and motivating our approach, however when the system is capable, this demonstrative example shows our method identifying that no human intervention is needed and autonomous manipulation proceeds successfully. While in this case the spatial uncertainty maps still identify regions with the highest influence over uncertainty, the agreement volatility plots show a relatively low overall agreement volatility. This indicates high ensemble confidence, allowing the system to complete the soft tissue manipulation task without requesting a human intervention.

Case 2: Intervention and Recovery In contrast, the second example (bottom row of Fig. 9) demonstrates an out-of-distribution scenario where the manipulation point is poorly selected, leading to a high chance of task failure. In this case, the agreement volatility plots show a relatively large overall initial agreement volatility particularly for the positional component. The spatial uncertainty map also immediately identifies the high agreement volatility near the manipulation point as well as the non-overlapping geometry of the goal point cloud. As a result, VAD-Net triggers a human intervention. After the human corrects the manipulation point, we see a steep decline in both the

positional and rotational agreement volatilities, indicating an increase in model confidence. VAD-Net subsequently reclaims control and successfully completes the task autonomously. This example highlights VAD-Net's ability not only to detect uncertainty in real-time, but also to attribute that uncertainty to regions of the input and reclaim control after human intervention.

Together, these cases demonstrate the value of agreement volatility not only as a signal of model uncertainty but also a tool for interpretability, enabling users to understand the source of uncertainty, which is critical for high-stakes applications like surgical robotics.