SeeSaw: Learning Soft Tissue Deformation From Laparoscopy Videos With GNNs

Reuben Docea, Jinjing Xu, Wei Ling, Alexander C. Jenke, Fiona R. Kolbinger, Marius Distler, Carina Riediger, Jürgen Weitz, Stefanie Speidel, and Micha Pfeiffer

***Abstract*—A major challenge in image-guided laparoscopic surgery is that structures of interest often deform and go, even if only momentarily, out of view. Methods which rely on having an up-to-date impression of those structures, such as registration or localisation, are undermined in these circumstances. This is particularly true for soft-tissue structures that continually change shape - in registration, they must often be re-mapped. Furthermore, methods which require 'revisiting' of previously seen areas cannot in principle function reliably in dynamic contexts, drastically weakening their uptake in the operating room. We present a novel approach for learning to estimate the deformed states of previously seen soft tissue surfaces from currently observable regions, using a combined approach that includes a Graph Neural Network (GNN). The training data is based on stereo laparoscopic surgery videos, generated semi-automatically with minimal labelling effort. Trackable segments are first identified using a feature detection algorithm, from which surface meshes are produced using depth estimation and delaunay triangulation. We show the method can predict the displacements of previously visible soft tissue structures connected to currently visible regions with observed displacements, both on patient data and porcine data. Our innovative approach learns to compensate non-rigidity in abdominal endoscopic scenes directly from stereo laparoscopic videos through targeting a new problem formulation, and stands to benefit a variety of target applications in dynamic environments.**

***Index Terms*—Deformation estimation, graph neural network (GNN), laparoscopy, machine learning, minimally invasive surgery, soft tissue deformation, surgical navigation.**

Manuscript received 19 January 2024; revised 17 May 2024; accepted 24 June 2024. Date of publication 15 July 2024; date of current version 22 November 2024. This work was supported in part by the State of Saxony through Sächsische Aufbaubank (SAB) in the scope of the ARAILIS Project under Grant 100400076. This measure is co-financed with tax funds on the basis of the budget passed by the Saxon state parliament. This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement Number 101092646, in part by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy - EXC 2050/1 under Project 390696704 and in part by the Cluster of Excellence Centre for Tactile Internet with Human-in-the-Loop (CeTI) as well as by the German Federal Ministry of Health (BMG) within the SurgOmics-project under Grant BMG 2520DAT82. *(Corresponding author: Reuben Docea.)*

Reuben Docea is with the National Center for Tumor Diseases (NCT/UCC Dresden), 01307 Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany (e-mail: reuben.docea@nct-dresden.de).

Jinjing Xu, Wei Ling, Alexander C. Jenke, and Micha Pfeiffer are with the National Center for Tumor Diseases (NCT/UCC Dresden), Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany.

Fiona R. Kolbinger is with the National Center for Tumor Diseases (NCT/UCC Dresden), Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany, and with the Else Kroener Fresenius Center for Digital Health, TUD Dresden University of Technology, Germany, and also with the Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Germany.

Marius Distler and Carina Riediger are with the Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Germany.

Jürgen Weitz is with the Department of Visceral, Thoracic and Vascular Surgery, University Hospital and Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Germany, and also with the Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden University of Technology, Germany.

Stefanie Speidel is with the National Center for Tumor Diseases (NCT/UCC Dresden), Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany, and with the Else Kroener Fresenius Center for Digital Health, TUD Dresden University of Technology, Germany, and also with the Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden University of Technology, Germany.

Project page for this work: https://gitlab.com/nct_tso_public/seesaw-soft-tissue-deformation

Digital Object Identifier 10.1109/TBME.2024.3424771

I. Introduction

COMPUTER Assisted Surgery (CAS) encompasses a large variety of techniques which aid surgical personnel in the carrying out of surgery. Many of these methods require an intraoperative model of the surgical site, such as registration with preoperative data [1], [2], [3], tracking areas of interest [4], [5], [6], [7] or for path planning in robotic assisted surgery [8], [9]. Keeping such a model up-to-date is a challenging task, especially in minimally invasive surgery (MIS, where the view is very limited), whenever soft-tissue is involved (which deforms due to breathing, heart beat and manual manipulation) or when smoke and tools obscure the surgical site from view. Furthermore, structures can deform when they are out of view [10]. In these cases, feature-tracking or mapping systems can lose track of parts or all of the surface [11]. In Simultaneous Localisation and Mapping (SLAM), recovery from such a failure after the scene has deformed can be extremely difficult.

To tackle these problems, the prediction of the current states of previously visible or hidden structures has been subject to research in recent years. Modelling the priors necessary to

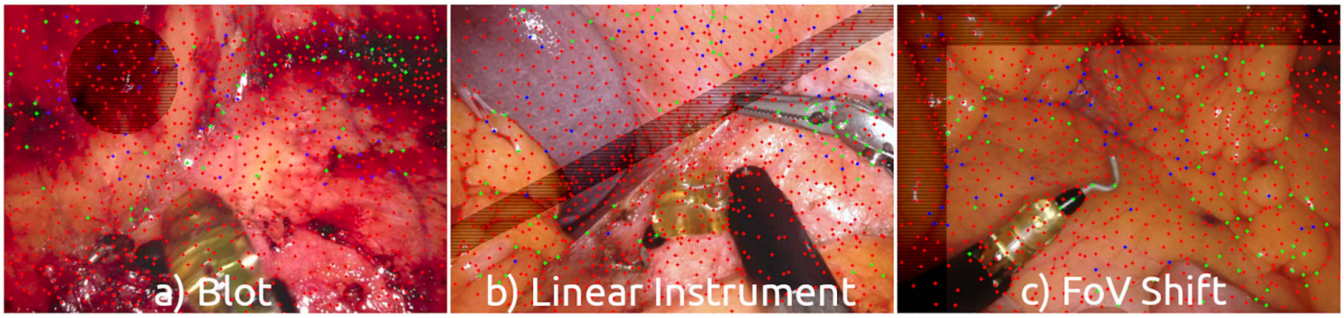

Fig. 1. Typical images from video sequences used to train our model. Instruments can obstruct the view of the tissue or the camera moves and tissue is no longer visible. Furthermore, pieces of tissue can move in different directions. Blue keypoints within dark regions are those which *can* be used for learning. However, for each particular case, only blue points which lie in masked regions (hatched regions in the images of this figure) are used to calculate a loss. Green and Blue points outside the dark regions are visible points input to the model. Red keypoints are points unmatched across timepoints. To reflect the different ways structures can be obscured from view, our creation of masks in the data generation process encompasses the concepts of interpolation and extrapolation. **a)** Represents a case where a blot on the camera obscures keypoints from view, **b)** reflects the case where an instrument occludes a portion of the tissue from view, and lastly, **c)** imitates the shifting of the camera such that the field of view doesn't capture previously seen tissue. This masking is done in a random manner.

do this for soft tissue is very challenging [12]. Relatedly, the non-rigid reconstruction of dynamic and deforming soft tissue has been done through modelling with neural fields [13], [14] or Gaussian splatting [15]. Despite their reconstruction accuracy, these methods are trained scene-wise and encode temporal deformations implicitly in the trained data. In other words, they do not learn how soft tissue deforms, but rather they learn the deformation of a particular scene. Therefore they can't be used for zero-shot online prediction of tissue deformation out of the distribution or across scenes. In contrast, the method we introduce learns to predict displacements, and therefore deformations, of previously seen soft tissue structures of the scene and generalises across datasets without fine tuning. More generally, there exists work on the non-rigid reconstruction of isolated subjects. Li et al. [16] released 4DComplete, a method for estimating the motion of occluded regions of moving objects. Lin et al. introduce OcclusionFusion [17], a dynamic 3D reconstruction method which advances on 4DComplete to perform in a real-time setting. Efforts towards shape completion, such as in the domain of Point Cloud Completion [18], [19], [20], also align well with our task. However, they are often limited to predefined classes of objects. The case is similar for Graph Convolutional Autoencoder-based methods [21], [22]. Sanchez-Gonzalez et al. showed that their Graph Network-based Simulators (GNS) can accurately simulate the behaviours of several complex materials [23]. Similarly, Pfaff et al. showed the ability of Graph Neural Network (GNN) methods to simulate mesh-based structures, including thin cloth [24]. Salehi & Giannacopoulos introduced 'PhysGNN' for estimating deformations of brain tissue under specific forces, learned from Finite Element Method (FEM) simulations of deformed brain meshes [25].

In estimating soft tissue deformation during laparoscopy, learning class-level models is inappropriate, and simulating the human abdominal environment is prohibitively complex. Instead, we present a novel two-stage method featuring a GNN which predicts, with respect to the current camera perspective, the up-to-date positions of previously seen soft tissue regions which are no longer visible, from partial views of these structures and while accounting for non-rigid changes realistically. In the scope of this work, we aim to do so for short-range deformations (less than 10 mm), which already stands to benefit downstream methods and applications such as tracking and SLAM. The method firstly roughly aligns the contents of the surgical scene at two different timepoints using feature matching and Singular Value Decomposition (SVD), followed by finely and non-rigidly compensating the residuals from SVD in a learned realistic manner using a GNN. Our method takes inspiration from Pfaff et al. [24], but rather than tackling physics simulation, it estimates the complete current state given the complete initial state and partial knowledge of the current state. We bypass the need for simulated data, which is difficult to obtain due to complexity and therefore laboriousness, by leveraging feature matching and depth estimation on video data. Training data is generated from stereo laparoscopic videos of abdominal surgeries (see Fig. 1), allowing our method to learn a pseudo-physical model for the prediction of deformations of abdominal soft tissue, directly from real data.

II. Methodology

The main novelty in our method is the learning process which allows us to train our GNN to realistically predict changes in the scene, including soft-tissue deformation. Although SVD alignment is an integral part of our overall approach, it is a fixed component. As such, we focus on details regarding GNN training and implementation in the methodology. We likewise focus on the GNN in evaluation, while assessing the contribution of SVD when appropriate. This study was performed in accordance with the ethical standards of the Helsinki declaration and its later amendments. The local Institutional Review Board (ethics committee at the Dresden University of Technology) reviewed and approved this study (approval number: BO-EK-140032021) on the 16th of May 2021. As this is a retrospective study on routine clinical data, written informed consent was not required. The trial was registered on clinicaltrials.gov (trial registration ID: NCT05268432).

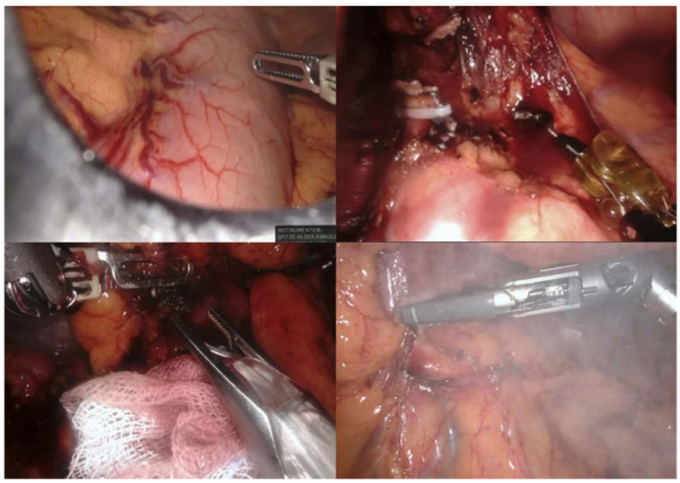

Fig. 2. Exclusion Criteria Examples. *Top Left:* Trocar and DaVinci UI elements visible. *Top Right:* Excessive blood pooling. *Bottom Left:* Objects other than instruments present (gauze). *Bottom Right:* Significant presence of smoke.

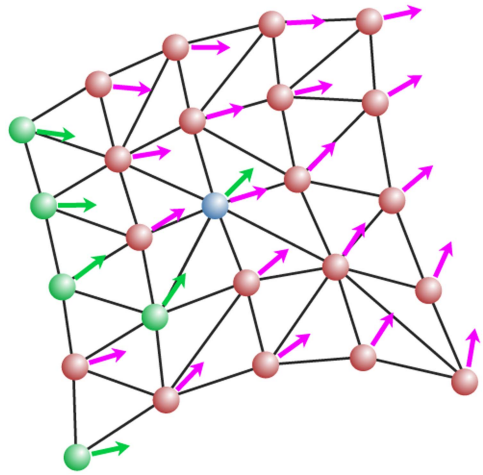

Fig. 3. Illustration of learning problem, based on Graph Process in Fig. 6. Part of the set of known visible residual displacements is hidden U_{hidden} (blue node), causing a prediction to be made for this node. The error between the predicted (magenta) and known residual displacements (green) U_{vis} for such nodes is used to train the GNN. Refer to Fig. 6 for diagram key.

TABLE I
NUMBER OF SEGMENTS PER VIDEO INDEX

Video Index	1	2	3	4	5	6	7	8	9	10	11	12
No. Segments	57	135	24	121	9	133	124	52	78	55	91	60

A. Data Generation

1) Extraction of Trackable Video Segments: First, we collect segments of video footage where the same structures are in motion or deforming - these are *trackable* segments. To do this, the feature detection algorithm SuperPoint [26] and the feature matching algorithm SuperGlue [27] were employed. We used abdominal laparoscopic surgery videos of 12 patients collected at the University Hospital Carl Gustav Carus Dresden, which are not publicly available, recorded from a da Vinci surgical robot with a stereo laparoscope. They ranged in length from 0.16 to 9.12 h (average 4.91 h, median 5.67 h). Across all videos, 8854 trackable segments were identified. Not all segments were suitable for learning soft tissue deformation, and so were filtered out according to exclusion criteria such as excessive complexity, limited soft tissue visibility or substantial pools of fluid (see Fig. 2). From this pool, 939 filtered segments were obtained in total, rejecting roughly 90% of examined segments. A break down of the number of segments per video can be seen in Table I, where their lengths have a mean and median of 5.5 s $\pm$ 4.8 s and 4.2 s, respectively.

2) Training Data Setup From Scene-State-Pairs: Using a random stereo frame of a segment at time t_p, we form a mesh M. Nodes of the mesh, V, are acquired by projecting keypoints, K_{t_p}, detected in the left image at t_p into 3D space with depth estimated using RAFT-Stereo [28]. The number of nodes in the mesh is increased to 1000 through Farthest Point Sampling (FPS) of a depth-map-derived pointcloud. This is done so that the graph captures more geometric information relating to the scene. Delaunay triangulation of V gives mesh edges E_{mesh}. To increase the reach of information, a subset of 121 of the final 1000 nodes in the graph are again sub-selected using FPS and fully connected to give supplementary edges, E_{supp}. The full set of edges, E, is the combination of the sets E_{mesh} and E_{supp}.

From matches between image keypoints at t_p and any other random frame at t_q, and a similar 3D-projection of keypoints K_{t_q} we obtain a set of primary displacements, $\mathbf{d} \in D_{match}$. Nodes which correspond to matches across timepoints t_p and t_q, we term V_{match}. We formulate our learning problem by artificially partitioning V_{match} into visible, V_{vis}, and hidden, V_{hidden}, learning to predict displacements on hidden nodes. This is visualised in Fig. 3. In order to maintain an up-to-date impression of previously visible soft tissue structures from the current view, it is necessary to estimate these primary displacements for those previously visible structures. We consider these primary displacements to be composed of a rigid component (for example, but not limited to, camera movement), $\mathbf{R}$, and a non-rigid component (soft tissue deformation or otherwise), $\mathbf{NR}$. To account for the rigid component $\mathbf{R}$, and to 'standardise' the learning problem for the GNN, we rigidly align the keypoints separated by the visible subset of D_{match}, D_{vis}, using Singular Value Decomposition (SVD).[1] Following SVD alignment, we are left with residual displacements, $\mathbf{u} \in U_{match}$. These residual displacements represent $\mathbf{NR}$ in our problem setup, and are to be compensated by the GNN. However, this is not to say that SVD does not partly account for displacements arising from deformation. In our learning setup, we thus learn to predict hidden residual displacements U_{hidden} from visible residual displacements U_{vis}.

Before partitioning, we first identify residual displacements that are likely to be well-estimated, or *trustworthy*. Such residual displacement values are used to backpropagate errors and train

[1][Online]. Available: https://github.com/nghiaho12/rigid_transform_3D

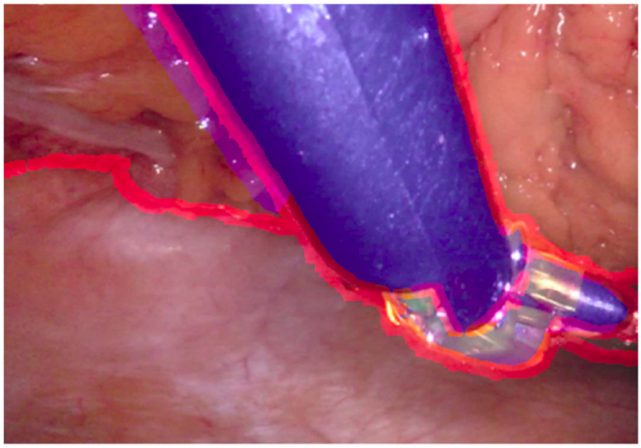

Fig. 4. Example image from video, displaying instrument mask (blue) and dilated depth-derived mask identifying regions with high gradients in the depth map (red). Both are used to identify nodes which should not contribute to the loss.

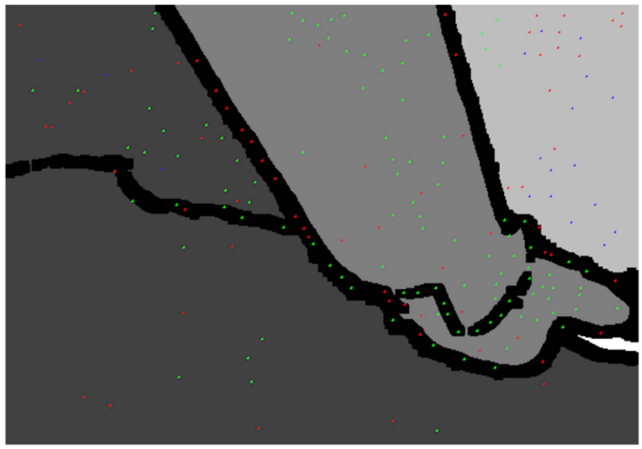

Fig. 5. Illustration of contiguous regions in depth map (indicated by differing shades of grey), with partitioned nodes V_{match}. Green nodes are visible, red nodes are invisible, and blue nodes are trustworthy. Nodes from FPS are not displayed here.

our GNN, with the expectation that they are not outliers. This is done by exempting nodes located on large gradients in the depth map, those found to belong to instruments through instrument segmentation [29] (see Fig. 4), as well as those matched by SuperGlue [27] with a confidence below a threshold. Nodes which remain are considered trustworthy. We then partition the points using three different types of mask, encompassing the concepts of interpolation and extrapolation (see Fig. 1):

- Blot (*interpolation*): a circular region centred on a random visible keypoint.
- Linear Instrument (*interpolation*): linear segment intersecting on a random visible keypoint, with random angle.
- Field of View Shift (*extrapolation*): edge of image offset from random corner is occluded.

The pixel-wise radius, width or displacement of the different mask elements are adjusted such that 80 to 85% of nodes in V_{match} fall outside the masked elements. Constrained by the following criteria, these nodes are taken to be the visible nodes, V_{vis} (see Fig. 5):

- *At least three trustworthy nodes must be allocated to* V_{hidden}. This is in an effort to mitigate the impact of learning from an outlier displacement value by increasing the chance non-outlying values are also present.
- *At least 3 matched nodes should be allocated to each contiguous region of the depth map where there is also at least one trustworthy node*. This is to account for the likely case that different contiguous regions will move more independently of one another: without information pertaining to a certain region, when the GNN estimates residual displacements of unseen nodes of that region, backpropagating errors from *trustworthy* residual displacements has the potential to be similar to the GNN learning from noise.

B. Graph Neural Network

To estimate the non-rigid component, **NR**, of our problem, we build a simple GNN framework (illustrated in Fig. 6), inspired by that in Pfaff et al. [24].

The nodes, V, of graph G have the following information: x, y, z primary and residual displacement vectors (zero-valued if node is not visible), as well as a Boolean value indicating if the node belongs to V_{vis}. This information is then encoded to a 128-dimensional feature vector using a 2-layer neural network block, ϵ_V, with $tanh$ activations and a skip-connection between first and last layers.

The edges, E, of graph G, are bidirectional and have the following information: the difference in position between the nodes they connect, in x, y, z coordinates, as well as the Euclidean distance between them. A separate encoder, ϵ_E, identical to ϵ_V in all but the input dimension, is used to encode information on edges.

Updates to the nodes and edges are then performed as follows, and in the following sequence, for L *message-passing* steps:

$$\mathbf{e}_{ij} \leftarrow f_E\left(\mathbf{e}_{ij}, \mathbf{v}_i, \mathbf{v}_j\right) \tag{1}$$

$$\mathbf{v}_i \leftarrow \tanh(\max_j(\mathbf{e}_{ij})) \tag{2}$$

Where f_E is again a neural network with a skip connection, i and j are node indices $i, j \in V$, $\mathbf{e}_{ij}$ is the edge connecting nodes i and j, and $\mathbf{v}_i$ and $\mathbf{v}_j$ are the node features for nodes i and j respectively.

To obtain the predicted residual displacements on V_{hidden}, another neural network, δ_V, lastly decodes $\mathbf{v} \in V$ to 3-dimensional vectors, i.e. estimated residual displacements $\hat{\mathbf{u}}$.

C. Model Training

The GNN was implemented in PyTorch, and trained in a 6-fold cross-validation (CV) on approximately 100,000 meshes, half extrapolation and half interpolation, across the 12 videos: 8 training, 2 validation, 2 testing. We refer to this as the 12-Patient dataset. The models were trained for 150 epochs on each fold, with a decaying learning rate from 10^{-4} to 10^{-6}. Training was carried out on RTX A5000 GPUs, requiring around 2.5 days per

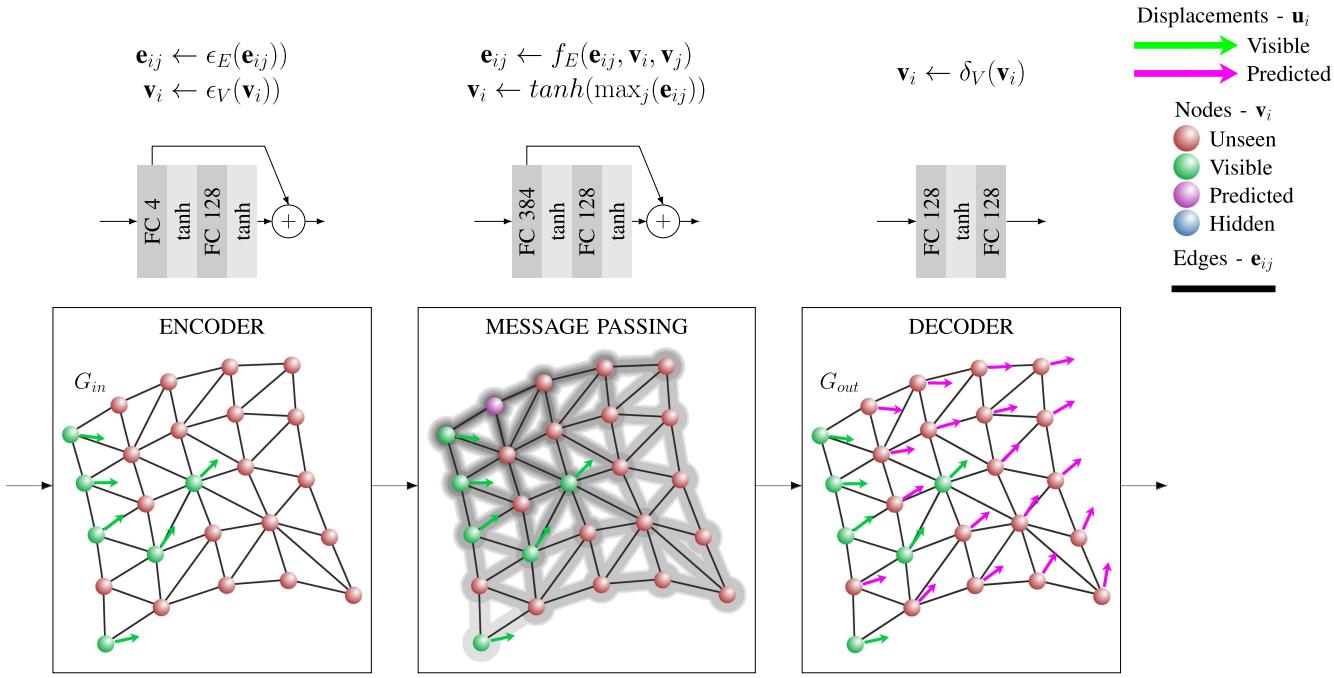

Fig. 6. Diagram illustrating the GNN-based approach introduced in this work. A graph, G_{in}, is received as input with a sparse set of visible displacements. To this graph, a 'Graph Process' is applied, which predicts the displacements of nodes which have no known displacements to produce the output graph, G_{out}. The diagram shows this Graph Process, with the structure of the different neural network blocks used, and a visualisation of a single iteration. The grey shading of 'Message Passing' illustrates the information reach of five steps of message passing centered on the purple node, with later steps of message passing corresponding to lighter greys.

fold. Latent feature dimensions for all neural network blocks was 128, each had 2 layers, and there were 5 message-passing steps. An L1 loss was used, and any nodes whose target residual displacement was greater than 3 s.d. above $||\bar{U}||_{match}$ (the mean euclidean norm of all residual displacements U_{match} in the current graph) were excluded from the loss calculation to further suppress outliers. The GNN runs at 100 Hz per mesh on a NVIDIA RTX A5000.

III. Evaluation

In our evaluation, we analyse both the second stage of the method which our GNN forms, as well as the performance of the complete SVD-GNN combination. We carry out this evaluation in terms of compensating residual displacements, U, at the level of the second-stage methods (GNN and otherwise), and ultimately evaluate the method as a whole with regards to the primary displacements, D. We assess the performance of our method and the presence of errors in our data generation process with a set of experiments:

1) Firstly, we investigate the errors relating to depth estimation, which is essential to the training and evaluation of our method. We inform the rest of our evaluation using findings here. We first assess the error arising from the RAFT-Stereo disparity matching algorithm in endoscopic abdominal scenes using the SCARED dataset [30]. We additionally gauge the error from depth estimation in our generated training dataset by assessing the estimated length of a tool with known dimensions in our dataset, the Cadiere Forceps.
2) We run the models of each CV fold on their corresponding test sets and amalgamated the results, which were overall similar across folds.
3) We evaluate the method using a different dataset from that which it was trained - the StereoMIS dataset [31].
4) Lastly, we qualitatively inspect the outputs of the model.

Due to the lack of comparable methods, we include two baselines. Firstly, a naive baseline method, which we refer to as Neighbour Averaging (NA), averages displacements of adjacent V_{vis} graph nodes to provide estimates for neighbouring nodes in V_{unseen}. This is done iteratively to propagate predictions throughout the graph structure. A second method we refer to as Gaussian-Weighted Averaging (GA), similar to the first, uses a Gaussian kernel to weight the influence of the visible displacements on the predictions on the unseen nodes. Here, the influence is a function of the distance of the node to be predicted for from the currently visible node. The influence is a function of the distance, computed with the following equation:

$$q_{j,i} = e^{-\frac{\|p_i - p_j\|_2^2}{2\sigma^2}} \tag{3}$$

where $q_{j,i}$ is the influence of node j on node i, and p_i is the position of node i. Lastly, σ is a constant value for the euclidean distance which corresponds to one standard deviation. The value of the constant for each test fold is determined experimentally using the Optuna optimisation framework [32] from 25 trials

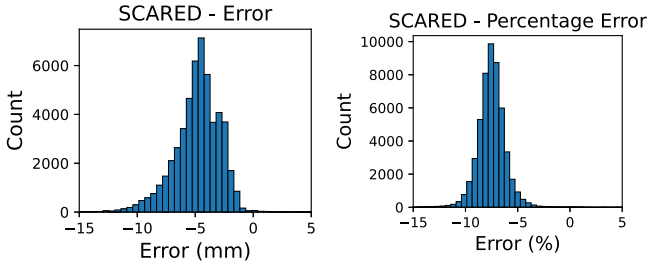

Fig. 7. Distribution of depth errors of RAFT-Stereo in evaluation on SCARED dataset. Mean error -4.85 mm $\pm$ 2.98 mm, or -7.46% $\pm$ 3.72%.

Fig. 8. Image of the da Vinci Cadiere Forceps, with the lumen length dimension (as we name it) indicated with a red arrow.

TABLE II
CADIERE FORCEPS SEGMENTS

Video	1	2	3	4	5	6	7	8	9	10	11	12
# Segments	2	3	0	5	0	6	4	3	4	1	1	1

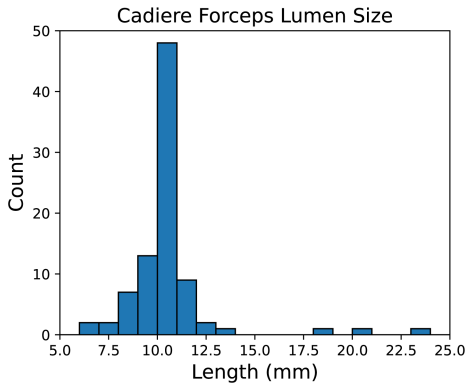

Fig. 9. Histogram of the depth-estimated lengths of the da Vinci Cadiere Forceps lumen, where the length of the lumen as measured with calipers is 12.3 mm.

and a random sample of approximately 8,000 graphs (10%) of the combined training and validation sets of each fold.

A. Data Generation Analysis

As the training and evaluation data in this work is generated from surgical videos, its accuracy is contingent on the algorithms employed and thus also contains errors. Here, we carry out an assessment of the depth estimation process, which we consider most likely to be the main source of positional inaccuracy in the data. To do so, we firstly evaluate the RAFT-Stereo depth estimation algorithm, which we have employed, on the SCARED endoscopic depth estimation dataset [30]. Secondly, to get an impression of the accuracy in the case of our own dataset, we determine the error in the measured dimensions of the Cadiere Forceps [33] (see Fig. 8) tools of the DaVinci surgical robot system, which is found throughout our training data.

1) Evaluation on SCARED: The SCARED dataset [30] contains 9 sub-datasets acquired from pigs using the da Vinci Xi system. A structured light illuminator is used in combination with the stereo endoscope to produce depth information for five different keyframes in each sub-dataset. Each keyframe consists of a unique positioning of the stereo endoscope and the structured light illuminator, together with a set of camera calibration parameters, the stereo image pair (left and right resolutions of 1280x1024) and the corresponding structured-light-estimated depth map. Due to errors in the calibration parameters of the fourth and fifth sub-datasets, we omit these from our evaluation. In our data generation process (Section II-A) we utilise keypoints detected with SuperPoint, which themselves are more distinct features in the image. Due to this, we expect that the estimated depth for keypoints is likely to be better than for non-textured regions of the image. To reflect this, we detect keypoints in the frames of SCARED and provide results only for pixels corresponding to keypoints. The the median and mean absolute errors over all of the keypoints in the data were 4.64 mm (or 7.51%) and 4.97 mm $\pm$ 2.78 mm (or 7.68% $\pm$ 3.25%), respectively. Visualisations of the error distributions can be found in Fig. 7. With reference to the figure, the mean deviation is consistent and is -4.85 mm. As such, it appears that there is a systematic error in underestimation. The standard deviation is 2.98 mm, which we take to be the noise floor in the depth estimation with respect to our subsequent evaluations.

2) Cadiere Forceps Measurement Evaluation: We follow up the evaluation of RAFT-Stereo on the SCARED dataset by obtaining an estimate for the errors in our own complete depth estimation setup (resulting from the stereo camera calibration parameters and disparity estimation, together). To do so, we randomly sample segments from the entire pool of segments across the 12 videos in our dataset from Section II-A. If the Cadiere Forceps, and its entire lumen (see Fig. 8), were visible in the left frame, we randomly sample three frames within the segment in which the lumen is visible in its entirety. We do this until we obtain 30 segments across the videos altogether (see Table II for the segment numbers).

Altogether, this provided 89 left frames (for which there are stereo pairs), where one frame was dropped due to being mis-sampled. For all of the 89 left image frames, we mark the beginning and end points of the measurement in Fig. 8. We then compute their positions in 3D space using the corresponding estimated depth map, and obtain an estimate for the lumen length. To obtain a reference value, we measure the lumen of the Cadiere Forceps using a pair of calipers, and find its length to be 12.3 mm. Through this process, the median and mean depth-estimated lengths of the lumen were 10.6 mm and 11.0 mm $\pm$ 3.10 mm, respectively. The distribution of the estimated length can be seen in Fig. 9. The length of the lumen is on average 1.3 mm smaller than the true value, which, at 10.6%, roughly corresponds to

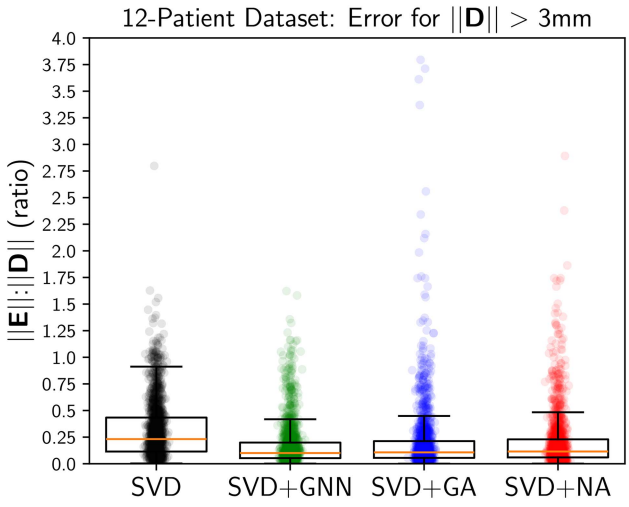

Fig. 10. Box plots illustrating the reduction in observed displacements greater than 3 mm (above depth estimation noise) for extrapolation and interpolation on our 12-Patient dataset, jointly. Results from SVD alone, and SVD combined with a subsequent stage (GNN, GA or NA) are shown. To reduce the size of the figure, only every 100th sample plotted.

a slightly greater under-estimation compared to the SCARED dataset.

B. Quantitative - Comparison on 12-Patient Dataset

Figs. 11 and 12 show results on the test sets. The error to the target position is well-minimised by both the baselines and our GNN when the distance to the nearest visible node is small. This is supported by the accumulation of points on the diagonal of the top left heatmap. When the size of the target residual displacement is at the larger end (closer to 10 mm), the GNN performs worse for $d(v) < 0.5$ cm. This is because the training data is dominated by smaller residual displacements (see Fig. 13). As the distance to the nearest visible node grows, the heatmaps become more dispersed as making accurate predictions increases in difficulty. For distances greater than 1.5 cm (bottom-right), the baseline methods in fact *increase* the distance to the target node position. On the other hand, with the exception of target residual displacements whose magnitudes are less than 1 mm, our method reduces the distance to the target position throughout almost all of the examined range. This is the case for both extrapolation and interpolation.

Ultimately, the purpose of the method is to estimate primary displacements, D. In this part of the evaluation, with reference to the noise floor in the depth estimation investigations, we consider only primary displacements, $\mathbf{d} \in D$, which are greater in magnitude than 3 mm. We present results for extrapolation and interpolation in Tables III and IV, where results are a ratio of the final error magnitude $||\mathbf{E}||$ over the primary displacement magnitudes $||\mathbf{D}||$ to account for differences in primary displacement magnitudes. We further illustrate these errors in Fig. 10. As can be seen, in eliminating the rigid component $\mathbf{R}$ of the D, SVD alone can provide an impression of the positions of previously seen structures. However, the combination of SVD with the GNN which we trained is uniformly both most effective, in terms of mean and median, and most reliable, in terms of standard deviation, across both extrapolation and interpolation splits.

TABLE III
12-PATIENT DATASET: REMAINING ERROR TO TARGET FOR EXTRAPOLATION ON **PRIMARY** DISPLACEMENTS GREATER THAN 3 mm

	Mean (%, (mm))	STD (%, (mm))	Median (%, (mm))
Primary	-, (5.08)	-, (2.66)	-, (4.17)
SVD	34.1, (1.54)	27.9, (1.36)	25.9, (1.17)
SVD+GNN	**19.1, (0.87)**	**21.9, (1.10)**	**11.2, (0.52)**
SVD+GA	21.4, (0.95)	39.3, (1.58)	11.7, (0.54)
SVD+NA	22.9, (1.03)	38.3, (1.72)	12.2, (0.57)

TABLE IV
12-PATIENT DATASET: REMAINING ERROR TO TARGET FOR INTERPOLATION ON **PRIMARY** DISPLACEMENTS GREATER THAN 3 mm

	Mean (%, (mm))	STD (%, (mm))	Median (%, (mm))
Primary	-, (5.14)	-, (2.67)	-, (4.27)
SVD	28.6, (1.30)	24.6, (1.24)	21.0, (0.96)
SVD+GNN	**15.2, (0.70)**	**17.8, (0.97)**	**9.2, (0.43)**
SVD+GA	17.8, (0.80)	38.0, (1.55)	9.9, (0.47)
SVD+NA	20.2, (0.91)	41.0, (1.71)	10.9, (0.51)

Comparing the computational demands, which may differ depending on implementation, the GNN is much faster than either baseline. It runs at 100 Hz in comparison to 45 Hz for the simple Neighbour Averaging, and 2 Hz for the more demanding Gaussian-Weighted Averaging. However, the NA and GA methods run on the CPU, and could be further optimised. Nevertheless, in its current form, the GNN presents a clear 'speed' advantage, making it much more promising for application to real-world real-time scenarios.

C. Quantitative - Evaluation on StereoMIS

In a similar way to the 12-Patient dataset, we perform an evaluation on a separate dataset - the StereoMIS dataset [31]. In this dataset, there are stereo endoscopic videos taken from three pigs using the da Vinci Xi surgical robot. We use the eight stereo sequences taken from the second pig, as all of the video data from the first pig contains structures which are not soft tissue, and data from the third pig is not publicly available. The images all have resolution 640x512. With this data, we construct an experiment to assess:

1) Whether our method works on different abdominal soft tissue data with a different set of camera parameters.
2) The extent to which our method accounts for soft tissue deformation.

For these purposes, we identify segments in the videos where the camera is still and where only soft tissue is present in the scene. To do so, we:

1) Identify periods in time for which there are surgical instruments in the scene that are completely still and occupy the same pixels in the camera image - S_{known}. Here, we can be certain that the camera is still because the da Vinci camera can only be moved separately from the tools: when the tools are perfectly still in the frame

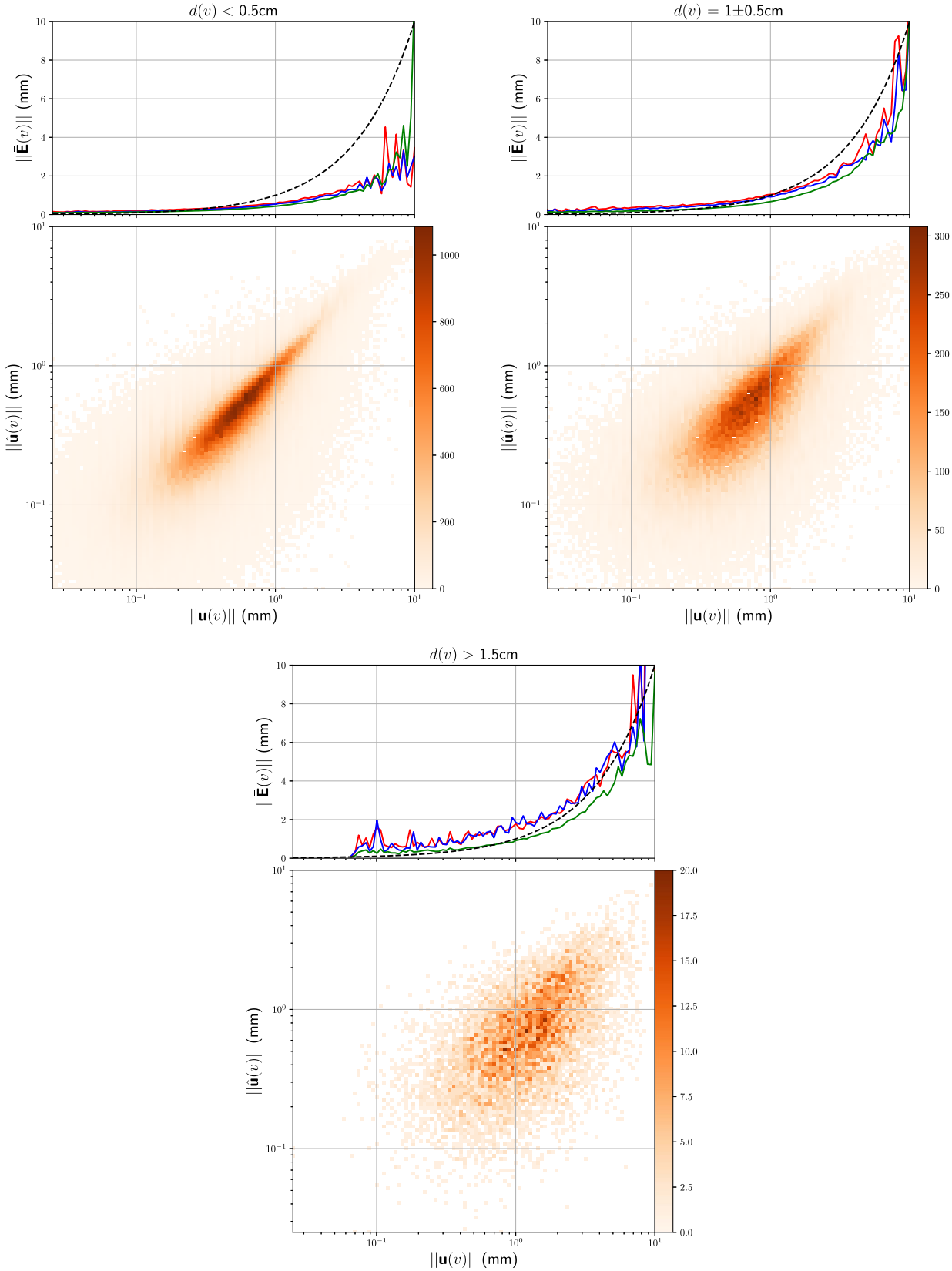

Fig. 11. Accuracy of GNN in predicting residual displacements for **extrapolation** (as illustrated in Fig. 1) at three different ranges of distance to nearest visible node, $d(v)$. The heatmaps compare the magnitude of the predicted displacement of a node v, $||\hat{\mathbf{u}}(v)||$, against that of the 'known' displacement, $||\mathbf{u}(v)||$, where colour intensity corresponds to the number of observations. Above each heatmap are plots comparing the average remaining distance to the target position of a node, i.e. the error, $||\bar{E}(v)||$, against the target displacement magnitude. Green lines represent GNN predictions, whereas red are those of NA, and blue of GA, and dashed lines indicate where the x- and y- axes are equal, i.e. no change in distance to the target position. Values below the dashed lines indicate improvements.

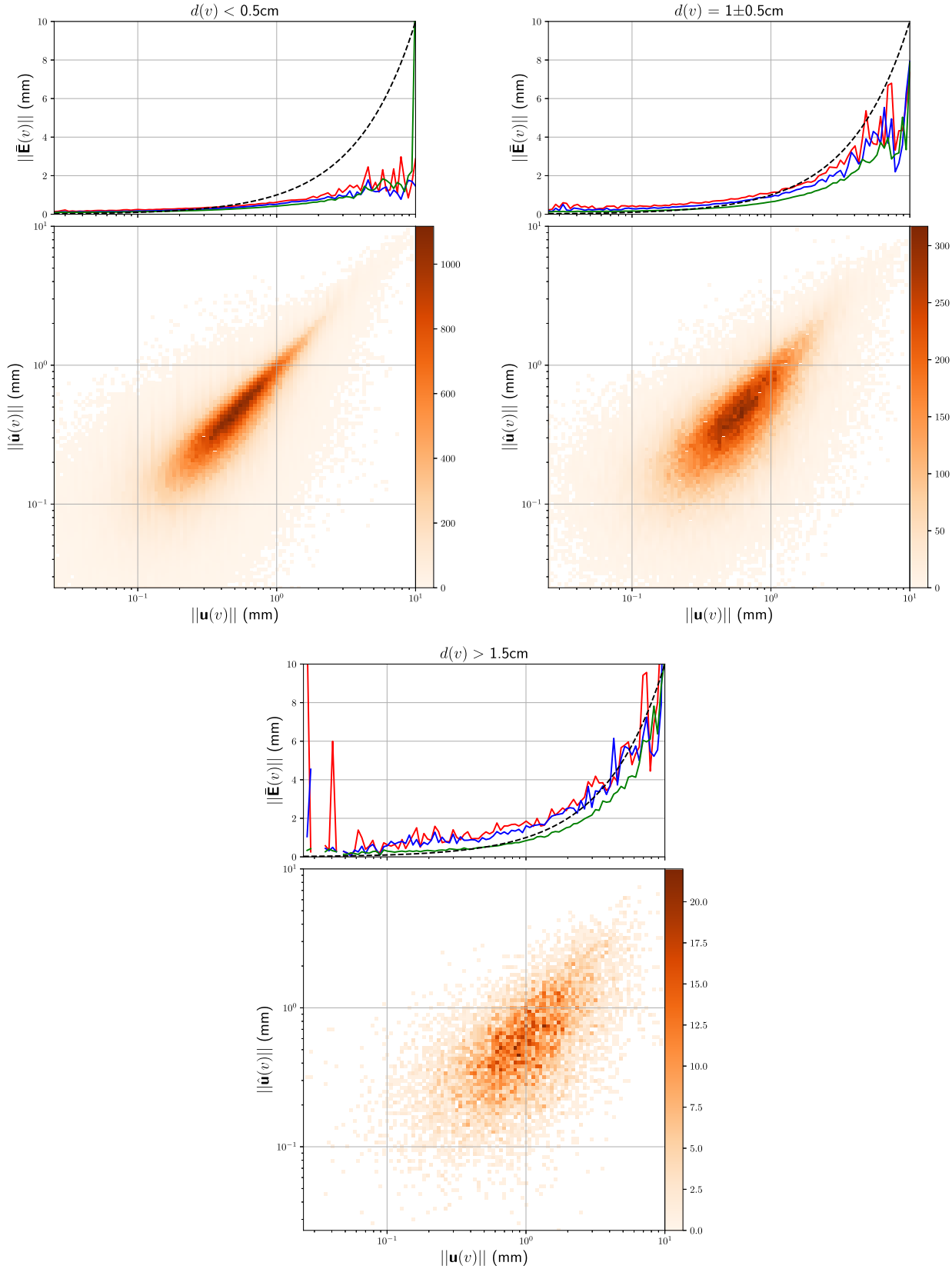

Fig. 12. Accuracy of GNN in predicting residual displacements for **interpolation** (as illustrated in Fig. 1), at three different ranges of distance to nearest visible node, $d(v)$. The heatmaps compare the magnitude of the predicted displacement of a node v, $||\hat{\mathbf{u}}(v)||$, against that of the 'known' displacement, $||\mathbf{u}(v)||$, where colour intensity corresponds to the number of observations. Above each heatmap are plots comparing the average remaining distance to the target position of a node, i.e. the error, $||\bar{\mathbf{E}}(v)||$, against the target displacement magnitude. Green lines represent GNN predictions, whereas red are those of the baseline method, and dashed lines indicate where the x- and y- axes are equal, i.e. no change in distance to the target position. Values below the dashed lines indicate improvements.

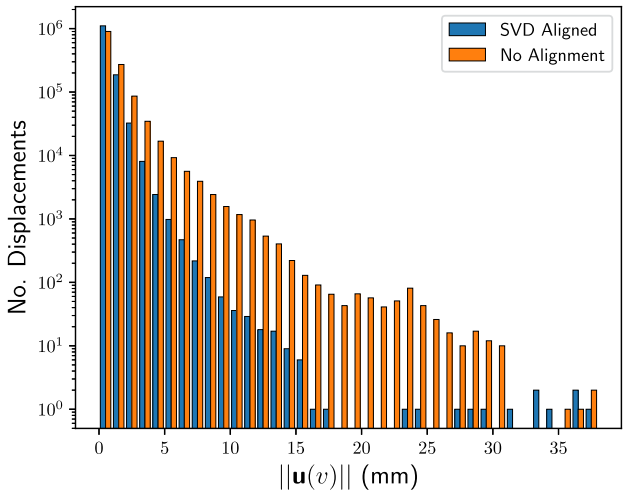

Fig. 13. Histogram plot showing the distribution of the subset of individual *trustworthy* displacement magnitudes, $||\mathbf{u}(v)||$, which were a part of V_{hidden}, throughout the entire dataset (all videos). The size distributions are shown for both before and after these are aligned by SVD.

the camera must also be still. This results in 10 segments across 4 videos, with a mean duration of 19.3 s $\pm$ 13.6 s.

2) With poses extracted from da Vinci kinematics which accompany the videos, we determine the standard deviation of the camera poses (in rotation and translation) for the pose sequences in S_{known}. We establish thresholds for stillness three times the size of the standard deviations: th_{rot} and th_{trans}.
3) Using th_{rot} and th_{trans}, we identify regions where the camera is still from periods in the pose sequences where the absolute camera pose did not change more than the threshold range, resulting in extracted segments $S_{extracted}$.
4) We visually filter the 61 segments obtained in $S_{extracted}$ to retain those where there is only soft tissue in the scene, and also apply the same criteria as in our earlier data generation step of Section II-A.
5) Due to our finding of inaccuracies in the kinematics pose sequences, we further visually sort through $S_{extracted}$, retaining only segments where the camera can be said to be static from visual inspection.
6) This gives us 33 segments across three videos. We further filter the segments, so that each ultimately retained segment represents a unique perspective from the endoscope. This is done by visually identifying the segments from the same perspective and randomly retaining one. Through this process, we are left with 11 segments across three videos, forming S_{still}. These segments are on average 5.11 s $\pm$ 4.47 s in length, with the shortest segment being 2.8 s in length.

We then proceed to use the video segments in S_{still} in the same data generation process as Section II-A. In this case, due to the higher resolution of the images, we increase the total number of nodes in the graph to 1650, through FPS, to be proportional with the number of pixels in the image. For each segment,

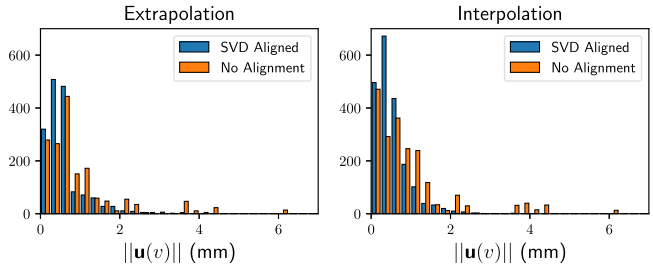

Fig. 14. Distribution of displacement magnitudes in StereoMIS.

TABLE V
STEREOMIS: REMAINING ERROR TO TARGET FOR EXTRAPOLATION ON **RESIDUAL** DISPLACEMENTS

	Mean (%, (mm))	STD (%, (mm))	Median (%, (mm))
Residual	-, (0.60)	-, (0.52)	-, (0.49)
GNN	**95.9, (0.57)**	**63.8, (0.54)**	**86.1, (0.39)**
GA	118.0, (0.81)	151.8, (2.11)	87.0, (0.51)
NA	139.3, (0.86)	349.4, (2.26)	93.3, (0.46)

TABLE VI
STEREOMIS: REMAINING ERROR TO TARGET FOR INTERPOLATION ON **RESIDUAL** DISPLACEMENTS

	Mean (%, (mm))	STD (%, (mm))	Median (%, (mm))
Residual	-, (0.53)	-, (0.42)	-, (0.43)
GNN	**88.5, (0.47)**	**51.8, (0.43)**	**82.5, (0.35)**
GA	118.5, (0.58)	161.9, (0.73)	88.2, (0.38)
NA	130.6, (0.66)	219.4, (0.94)	93.4, (0.41)

we generate **one** graph, G_s, between two random timepoints in the segment, $s \in S_{still}$. Differently from the standard data generation, for each graph, we inspect random matches provided by SuperGlue. We verify matches visually until we have 30 verified matches, M_s, per segment s. This leaves us with 330 visually verified matches, M_S, altogether. For each graph G_s, we generate 30 random extrapolation masks, and 30 random interpolation masks (each centred on a different matched keypoint in M_s). Altogether, this provides 660 different extrapolation and interpolation test graphs for the evaluation. Fig. 14 shows the distributions of the sizes of individual target displacements across the dataset based on M_S. Given that the camera is still, only soft tissue is visible, and the displacement distributions shift towards smaller values after SVD is applied, it is safe to say that SVD partly compensates for the observed soft tissue deformation.

As before, we evaluate the GNN, Gaussian-Weighted Averaging and Neighbour Averaging methods on the obtained graphs. Due to the smaller amount of available data for this test data, we summarise the results for all of the displacements where the target displacements are greater than 0.25 mm in size. We choose 0.25 mm as this is where the methods begin to make a difference as seen in Figs. 11 and 12. The remaining error after applying the three methods is reported as a ratio of the target displacement size to account for differences in displacement sizes. Results can be seen in Tables V and VI. Here, it can be seen that the GNN is

TABLE VII
STEREOMIS: REMAINING ERROR TO TARGET FOR EXTRAPOLATION ON **PRIMARY** DISPLACEMENTS GREATER THAN 3 mm

	Mean (%, (mm))	STD (%, (mm))	Median (%, (mm))
Primary	-, (4.19)	-, (0.82)	-, (3.91)
SVD	27.6, (1.20)	**22.2, (1.04)**	16.7, (0.62)
SVD+GNN	**22.1, (0.97)**	24.0, (1.03)	**8.2, (0.32)**
SVD+GA	52.5, (2.16)	118.8, (4.79)	8.4, (0.33)
SVD+NA	39.6, (1.73)	106.5, (4.62)	9.8, (0.39)

TABLE VIII
STEREOMIS: REMAINING ERROR TO TARGET FOR INTERPOLATION ON **PRIMARY** DISPLACEMENTS GREATER THAN 3 mm

	Mean (%, (mm))	STD (%, (mm))	Median (%, (mm))
Primary	-, (4.18)	-, (0.69)	-, (3.92)
SVD	17.2, (0.75)	**14.9, (0.72)**	13.3, (0.56)
SVD+GNN	**14.2, (0.63)**	17.0, (0.77)	9.1, (0.36)
SVD+GA	16.4, (0.70)	21.6, (0.92)	8.9, (0.37)
SVD+NA	17.7, (0.76)	43.6, (1.65)	**7.5, (0.30)**

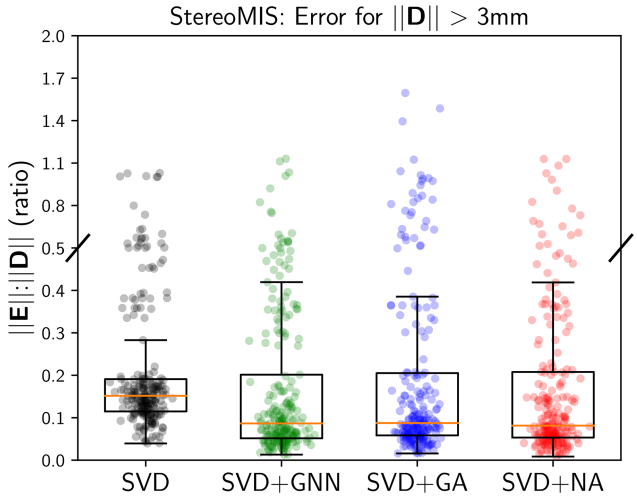

Fig. 15. Box plots illustrating the reduction in primary displacements greater than 3 mm (above depth estimation noise) for extrapolation and interpolation on StereoMIS, jointly. There are 233 displacements altogether (100 in extrapolation, 133 in interpolation). Results from SVD alone, and SVD combined with a subsequent stage (GNN, GA or NA) are shown.

most effective and reliable in estimating residual displacements U, throughout both extrapolation and interpolation.

However, as previously mentioned, the residual displacements U do not fully encompass deformation. In this experimental setup (fixed camera, soft tissue only), the primary displacements D serve this purpose. As in Section III-B, we further examine the estimation of primary displacements of magnitudes greater than 3 mm and therefore above the depth estimation noise floor. Results can be seen in Tables VII and VIII, and Fig. 15.

Except in terms of standard deviation, combining SVD with the GNN always results in an improvement over solely employing SVD. With reference to mean and std. dev. of Tables VII and VIII, our method combining SVD and GNN produces the greatest and most reliable prediction of primary displacements greater than 3 mm: with accuracies of 77.9% in extrapolation and 85.8% in interpolation.

D. Qualitative Results

Figs. 16 and 17 show example outputs from our network. In Fig. 16 several structures are visible - fatty tissue, and two other distinct regions which are under manipulation. From inspecting the displacements, our network clearly identifies the separate regions of the scene, as these all have displacement clusters that point in distinct directions. Furthermore, a number of outlying 'visible' displacements, which are highlighted by cyan arrows, are clearly and correctly ignored by the GNN in making its predictions. This demonstrates our network's robustness to incorrect feature matching. These are all testaments to the effectiveness of our data generation scheme, which has allowed our network to learn to discern the behaviour of soft tissues from highly complex scenes, learning to take geometric cues into account which may otherwise be difficult to do.

IV. DISCUSSION

Our assessments of RAFT-Stereo and the depth estimation processes in Sections III-A1 and III-A2 provide insight into the errors present in depth estimation, which our method relies on for obtaining ground truth data. Through these assessments, we find that depth estimation, as a composite of RAFT-Stereo and stereo camera calibration parameters, appears to slightly systematically underestimate the depth of points in the stereo endoscopic scene. To a large extent, it seems that this may be due to an underestimated baseline value, as when, in separate experiments, we scale the baseline, the errors are reduced by more than half across the board on the SCARED dataset. This suggests that this may be a problem aspect of stereo endoscope calibration more generally. Another limitation of our method is the dataset on which is was trained. Using surgeries on only 12 patients from one hospital leaves room for improvement to generalisation, where having a greater number of surgeries from a greater diversity of endoscopes and hospitals would deliver a benefit. Nevertheless, using the finding in our depth estimation investigations to inform our subsequent experiments and analyses allows us to concretely demonstrate the effectiveness of our method despite limitations in the training data. Furthermore, our application of these findings (a noise floor of 3 mm), to a separate dataset with a different endoscope and camera parameters, StereoMIS, evidences the generalisation capability of our method.

With reference to Figs. 11 and 12, our GNN is more capable of estimating displacements of regions further from visible nodes (more than 0.5 cm away) than the baseline methods. One reason for the worse baseline performances might be that they are more susceptible to outliers in U_{vis}. And a second reason is that errors are propagated throughout their predictions. On the other hand, our method consistently reduces the error to the target position for residual displacements above 1 mm in the range from 1 mm to 10 mm. Beyond this range, our network is less effective due to lacking training data.

To evaluate the primary displacements on the 12-Patient dataset, we limit the analysis to primary displacements which have a magnitude greater than 3 mm. Applying a second stage (i.e. GNN, GA or NA) after SVD is clearly beneficial to the task

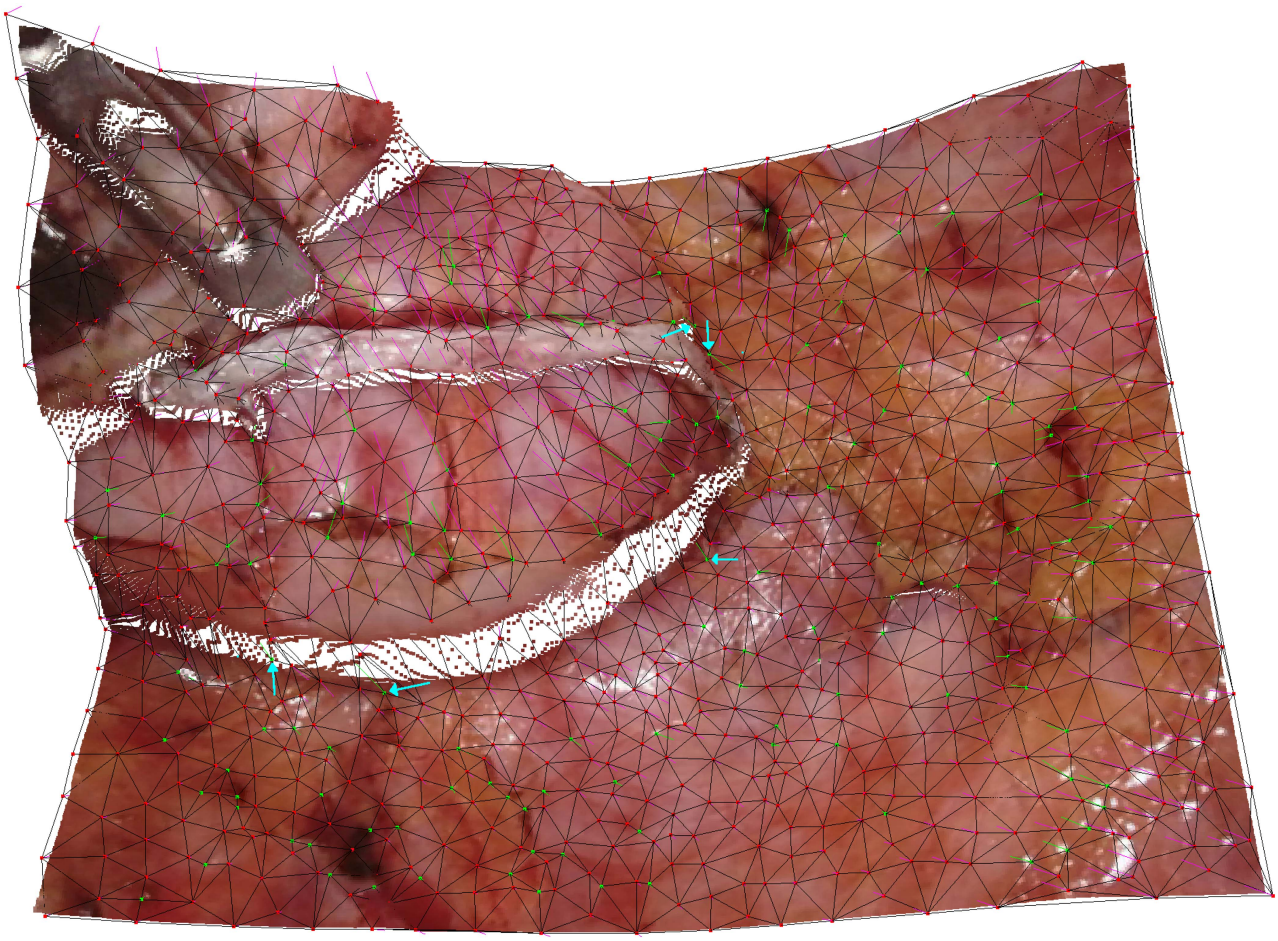

Fig. 16. Example graph structure output by our method, superimposed on pointcloud of scene. Inputs to the GNN are the green nodes, V_{vis}, and the green lines, U_{vis}. All red nodes are both invisible and untrustworthy. Purple lines are the predicted displacements output by our network. Cyan arrows point to visible displacements which are outliers, which our network successfully ignores. The mean and median sizes of input residual displacements are 1.37 mm $\pm$ 1.22 mm and 1.13 mm, and those of predicted residual displacements in the output are 1.09 mm $\pm$ 0.67 mm and 0.95 mm.

of predicting the displacement of previously seen structures. Furthermore, with reference to Tables III and IV, we find that combining SVD with our GNN is highly effective in both extrapolation and interpolation, and is more effective than any of the other combinations.

In our experimental design with the StereoMIS dataset, the endoscopic camera is still and only soft tissue is in view. We see that in the process of estimating the displacements of previously seen structures (composed of SVD & GNN together), the method remains highly effective in predicting primary displacements greater than 3 mm, roughly 77.9% accuracy in extrapolation and 85.8% accuracy in interpolation. These results correspond to our findings on the 12-Patient dataset in the Tables III and IV of Section III-B. As such, this experiment and the corresponding results conclusively prove the value of our method in compensating the deformation of soft tissue. Simultaneously, they demonstrate that SVD has a role in compensating deformation and restate the value of the approach we have taken, enabling us to learn the behaviour of soft tissue directly from real surgical videos.

Our qualitative evaluation concludes our evaluation by illustrating our method's ability to realistically predict displacements of structures in the surgical scene, and doing so all the while suppressing the influence of outliers in the input data.

The experiments and subsequent analysis that we have carried out clearly demonstrate our method's effectiveness in estimating the displacements of previously visible structures that fall out of view as a result of occlusion or otherwise. They show this through the performance on the scenarios of both extrapolation and interpolation, where, for short-range displacements, soft tissue deformation is realistically accounted for. With this, we see our method as having great potential for integration into a variety of more extended methods with downstream tasks in MIS. Firstly, our method could provide an inductive bias to tracking methods where a structure leaving view often terminates the tracking process, and enable the method to confidently resume tracking when structure returns to view. Furthermore, having a realistic prediction for how a tissue structure deforms may also enhance the tracking process for tracked structures that are currently in view.

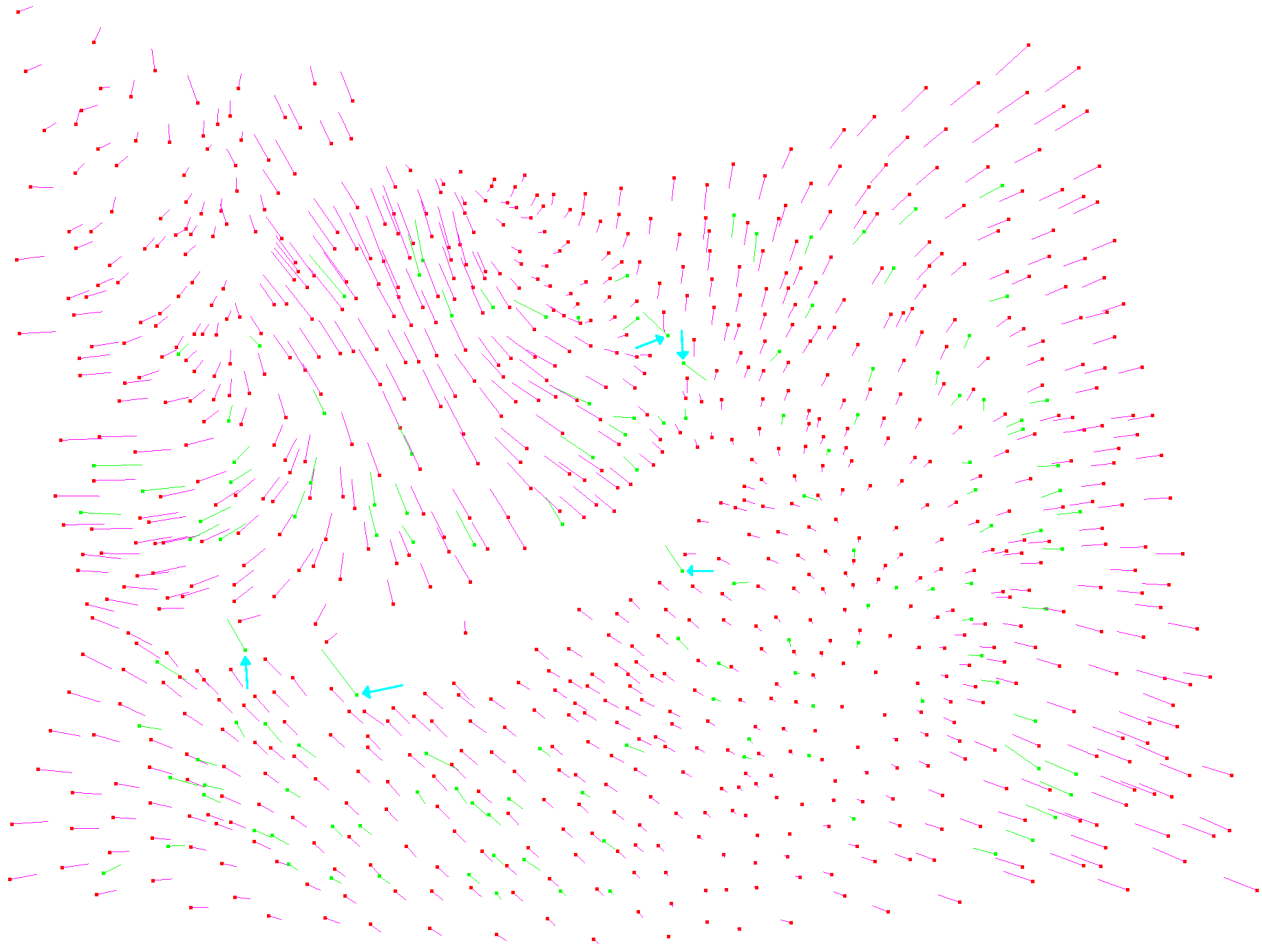

Fig. 17. Displacement field for the example in Fig. 16, but without the pointcloud and mesh. The network learns to predict the displacement of separate tissue regions from a subset of trackable displacements. Displacements and nodes are illustrated using same key as in Fig. 6. Cyan arrows point to visible displacements which are outliers, which our network successfully ignores.

A second task which the method's capabilities can certainly benefit is that of registration, where having a good current impression of organ surfaces, such as that of the liver, is crucial. In this vein, our method complements SLAM, wherein several works aim to tackle the challenge of deforming scenes, and real-time performance is a must. One such effort is that of MIS-SLAM [34], where an embedded deformation graph is used to accommodate soft tissue deformation in mapping, by minimising an energy function balancing uniformly-applied heuristics of how soft tissue is expected to deform. DefSLAM [35] and SD-DefSLAM [36] likewise address non-rigid SLAM in laparoscopy, however do not aim to estimate the current state of previously visible structures. Three properties of our work make it ideal for integration into SLAM systems. Firstly, as the method is based on feature detection, it can comfortably fit within the keyframe framework of feature-based SLAM. Secondly, the GNN method itself runs in a real-time fashion. And lastly, because the method uses whole-image data (not concentrating on isolated structures), and learns to handle the associated complexity of surgical scenes inherently, it is not limited to functioning on an isolated subject as with other learning-based methods such as 4DComplete [16] and OcclusionFusion [17]. A SLAM system integrating our method would likely resemble MIS-SLAM most closely, but with the main difference that rather than relying exclusively on heuristics to estimate map deformation, a method that is inherently sensitive to laparoscopic scene geometry and soft tissue deformation would feature in its place.

V. Conclusion

This work introduces a novel graph-based method for learning the behaviour of soft tissue directly from the highly complex environment of surgical videos, by leveraging advances in feature detection and depth estimation, and with only minimal manual input to dataset curation. Our method outperforms the baselines in both extrapolation and interpolation, throughout almost the entire dataset. With it clear that the network is capable of estimating changes in positions of nearby nodes, it is safe to say that the method can help to provide an updated view of previously visible areas. In its current form, it makes significant headway towards tackling the challenge of non-rigidity in contexts such as MIS, by standing to supplement or enhance the functioning of a range of methods such as liver registration, polyp tracking and SLAM, among other downstream tasks.

There remain promising avenues for further work, however. By maximising the variety of displacement sizes the training data contains (also greater than 10 mm) through improved data generation and a greater variety of training datasets, the performance of our method would likely improve. Furthermore, the inclusion of RGB values from images to the GNN input may benefit the method's accuracy through the inference of material properties. Likewise, including temporal information, through optical flow or otherwise, also stands to enhance the performance of the method. Future directions may also include development of a means to learn how to mesh the scene for best results. Learned re-meshing in this context could go further, making updates to the mesh connections upon cutting, by predicting the mesh structure that best explains observed displacements. By incorporating a time component, another possibility would be to integrate the prediction of regular motion of tissues caused by pulsation or breathing. With these in mind, in the near term, the integration of the method into a SLAM system will be investigated.

References

[1] M. Pfeiffer et al., "Non-rigid volume to surface registration using a data-driven biomechanical model," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, Cham, 2020, pp. 724–734.

[2] R. Docea et al., "Simultaneous localisation and mapping for laparoscopic liver navigation: A comparative evaluation study," *Proc. SPIE*, vol. 11598, pp. 62–76, 2021.

[3] R. Docea et al., "A laparoscopic liver navigation pipeline with minimal setup requirements," in *Proc. IEEE 2022 Biomed. Circuits Syst. Conf. (BioCAS)*, 2022, pp. 578–582.

[4] H. Zheng et al., "Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained CNN," in *Proc. IEEE 2019 16th Int. Symp. Biomed. Imag.*, 2019, pp. 79–82.

[5] H.-S. Tong et al., "Real-to-virtual domain transfer-based depth estimation for real-time 3D annotation in transnasal surgery: A study of annotation accuracy and stability," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, pp. 731–739, 2021.

[6] M. Carstens et al., "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," *Sci. Data*, vol. 10, no. 1, pp. 1–8, 2023.

[7] F. R. Kolbinger et al., "Anatomy segmentation in laparoscopic surgery: Comparison of machine learning and human expertise–An experimental study," *Int. J. Surg.*, vol. 109, no. 10, pp. 2962–2974, 2023.

[8] E. Tagliabue et al., "Intra-operative update of boundary conditions for patient-specific surgical simulation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, Berlin, Heidelberg, 2021, pp. 373–382.

[9] J. Lu et al., "Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction," in *Proc. IEEE 2021 Int. Conf. Robot. Automat. (ICRA)*, 2021, pp. 4783–4789.

[10] P. Azagra et al., "Endomapper dataset of complete calibrated endoscopy procedures," *Sci. Data*, vol. 10, no. 1, 2023, Art. no. 671.

[11] M. J. Fulton et al., "Comparing visual odometry systems in actively deforming simulated colon environments," in *Proc. 2020 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2020, pp. 4988–4995.

[12] J. Schüle et al., "A model-based simultaneous localization and mapping approach for deformable bodies," in *Proc. 2022 IEEE/ASME Int. Conf. Adv. Intell. Mechatron.*, 2022, pp. 607–612.

[13] Y. Wang et al., "Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2022, pp. 431–441.

[14] R. Zha et al., "EndoSurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2023, pp. 13–23.

[15] Y. Huang et al., "Endo-4DGS: Distilling depth ranking for endoscopic monocular scene reconstruction with 4D Gaussian splatting," 2024, *arXiv:2401.16416*.

[16] Y. Li et al., "4DComplete: Non-rigid motion estimation beyond the observable surface," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12706–12716.

[17] W. Lin et al., "OcclusionFusion: Occlusion-aware motion estimation for real-time dynamic 3D reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1736–1745.

[18] P. Xiang et al., "SnowflakeNet: Point cloud completion by snowflake point deconvolution with skip-transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5499–5509.

[19] W. Yuan et al., "PCN: Point completion network," in *Proc. 2018 Int. Conf. 3D Vis. (3DV)*, 2018, pp. 728–737.

[20] J. Wang et al., "PointAttN: You only need attention for point cloud completion," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 6, pp. 5472–5480, Mar. 2024.

[21] O. Litany et al., "Deformable shape completion with graph convolutional autoencoders," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1886–1895.

[22] S. Foti et al., "Intraoperative liver surface completion with graph convolutional VAE," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, Berlin, Germany: Springer, 2020, pp. 198–207.

[23] A. Sanchez-Gonzalez et al., "Learning to simulate complex physics with graph networks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8459–8468.

[24] T. Pfaff et al., "Learning mesh-based simulation with graph networks," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=roNqYL0_XP

[25] Y. Salehi and D. Giannacopoulos, "PhysGNN: A physics–driven graph neural network based model for predicting soft tissue deformation in image–guided neurosurgery," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 37282–37296. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/f200119a40846e508954abcd61f5f3fd-Paper-Conference.pdf

[26] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2018, pp. 337–33712.

[27] P.-E. Sarlin et al., "SuperGlue: Learning feature matching with graph neural networks," in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 4937–4946.

[28] L. Lipson, Z. Teed, and J. Deng, "RAFT-stereo: Multilevel recurrent field transforms for stereo matching," in *Proc. IEEE 2021 Int. Conf. 3D Vis. (3DV)*, 2021, pp. 218–227.

[29] A. C. Jenke et al., "Stay focused-enhancing model interpretability through guided feature training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 121–129.

[30] M. Allan et al., "Stereo correspondence and reconstruction of endoscopic data challenge," 2021, *arXiv:2101.01133*.

[31] M. Hayoz et al., "Learning how to robustly estimate camera pose in endoscopic videos," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, no. 7, pp. 1185–1192, 2023.

[32] T. Akiba et al., "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2623–2631.

[33] Intuitive Surgical, "Da Vinci XI & X system instruments and accessories catalog," 2023. Accessed: May 12, 2024. [Online]. Available: https://www.intuitive.com/en-us/-/media/ISI/Intuitive/Pdf/xi-x-ina-catalog-no-pricing-us-1052082.pdf

[34] J. Song et al., "MIS-SLAM: Real-time large-scale dense deformable SLAM system in minimal invasive surgery based on heterogeneous computing," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 4068–4075, Oct. 2018.

[35] J. Lamarca et al., "DefSLAM: Tracking and mapping of deforming scenes from monocular sequences," *IEEE Trans. Robot.*, vol. 37, no. 1, pp. 291–303, Feb. 2021.

[36] J. J. Gómez-Rodríguez et al., "SD-DefSLAM: Semi-direct monocular slam for deformable and intracorporeal scenes," in *Proc. IEEE 2021 Int. Conf. Robot. Automat. (ICRA)*, 2021, pp. 5170–5177.