International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-026-03583-6

ORIGINAL ARTICLE

Exploring video recognition models for force estimation in small bowel surgical retractions

Kevin Wang[1,2,3,4,5] · **Ariel Rodriguez**[1,2,3,4,6] · **Micha Pfeiffer**[1,2,3,4] · **Sebastian Bodenstedt**[1,2,3,4,5] · **Rayan Younis**[5,7] · **Martin Wagner**[5,7] · **Stefanie Speidel**[1,2,3,4,5,6]

Received: 13 August 2025 / Accepted: 9 February 2026

Abstract

Purpose: Excessive retraction force during a minimally invasive surgery could cause tissue damage, affecting the surgical outcome. The integration of surgical force feedback relies on incorporating force sensors into the instruments, predominantly for robot-assisted surgery. However, such hardware integration is costly and challenging, and such instruments for conventional laparoscopic surgery are not yet widely adopted. We propose to use video as the primary source to objectively estimate tissue retraction force, which presents a promising approach to assess the skill of surgeons while retracting bowel tissue and addressing the current haptic deficiency without the need for special instruments.

Methods: We first introduce an experimental setup for the acquisition of a force-vision dataset with conventional laparoscopic instruments retracting a silicone small bowel phantom as an example. A novel data collection procedure is described along with the corresponding force-signal preprocessing methods. Then, we explore the performance of state-of-the-art ResNet and transformer-based computer vision models for force estimation. Two experiments are undertaken to assess feasibility of vision-based force estimation models.

Results: Results show that the models generalize across the collected dataset with varying phantom geometries and camera angles. We find ResNet-based models outperforming transformers, with all results matching or surpassing previous works. We further show that all models could perform real-time force estimation with respect to our data collection rate, with 3D ResNet being the fastest.

Conclusion: Vision-based force estimation based on the laparoscopic video feed is a promising way toward force estimation of tissue retraction. A novel use of computer vision models has been proposed with a low-cost data collection pipeline. Since the proposed clinical usage does not require customized equipment, seamless integration to the existing surgical framework is possible. The experimental results additionally demonstrate potentials in the approach to pave the way for real-time haptic feedback and quantitative skill assessment.

Keywords Laparoscopic surgery · Robot assisted surgery · Haptics · Computer vision

Kevin Wang and Ariel Rodriguez contributed equally to this work.

✉ Kevin Wang
kevin.wang@nct-dresden.de

✉ Ariel Rodriguez
ariel.rodriguezjimenez@nct-dresden.de

[1] Translational Surgical Oncology, National Center for Tumor Diseases, 01307 Dresden, Germany

[2] German Cancer Research Center (DKFZ), Heidelberg, Germany

[3] Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany

[4] Faculty of Medicine and University Hospital Carl Gustav Carus, TUD, 01307 Dresden, Germany

[5] Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD, 01069 Dresden, Germany

[6] BMFTR Research Hub 6G-Life, TUD, 01069 Dresden, Germany

[7] Department of Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD, 01307 Dresden, Germany

Published online: 10 March 2026

Introduction

Gentle handling of tissues is considered to be the first fundamental principle of surgical skill and practice [1]. This holds true as excessive retraction force can lead to trauma in many organs, adversely affecting the quality of surgery [2], which in turn negatively influences postoperative outcomes of patients [3, 4]. In gastrointestinal surgery, the small intestine is often manipulated to provide better access to the operating field, but it is particularly vulnerable to injury from excessive retraction [5, 6]. It should be noted that small bowel retraction is commonly performed during various gastrointestinal laparoscopic cancer surgeries, such as ileocolostomy, to facilitate the lifting and dissection of the bowel under traction or to expose adjacent anatomical structures obscured by the bowel. At the beginning of one such surgery, bowel retraction often forms part of the exploratory phase, as the small intestine frequently overlies other critical anatomical regions. In non-oncological procedures, manipulation of the small bowel is also common, particularly during laparoscopic management of small bowel obstructions.

The minimally invasive technique is often used to perform gastrointestinal surgery [7], with conventional laparoscopic surgery being significantly more widespread than robot-assisted surgery (RAS). It consists of inserting thin rod-shaped instruments into the abdomen, but the limited force feedback provided by the instruments in conventional laparoscopic surgery, as well as the complete loss of force feedback in RAS, makes it challenging to prevent excessive retraction force [8]. Since the introduction of the da Vinci 5 Surgical System (Intuitive Inc., Sunnyvale, USA) with force sensors, the measurement of retraction force and even real-time force feedback are now possible for RAS. However, this benefit is still unavailable for the significantly more widespread conventional laparoscopic surgery and present RAS systems without custom-made mechanical sensors [9, 10]. In the meanwhile, it has been shown that experienced surgeons can deduce the force applied by their laparoscopic instruments using only visual cues [11], also referred to as 'visual-haptics.' As such, gentle tissue retraction in laparoscopic surgery is a skill that comes with experience [12, 13].

Thus, the measurement of retraction force after surgery could be helpful to assess surgical quality and to inform novice surgeons to improve their surgical skill [14]. However, it remains a challenge to quantify retraction force of the small bowel during laparoscopic procedures on a large scale. Popular surgical skill assessment tools include tissue handling as a component, but it remains a subjective measure [15, 16].

Following the approach of previous works leveraging computer vision (CV) on force feedback for RAS [17, 18], our CV methods—similar to surgeons using 'visual-haptics'—aim to address this deficiency—especially for laparoscopic procedures. By adopting vision-based force estimation models, we eliminate difficulties surrounding mechanical sensors. Hence, our methods could be applied to both RAS and laparoscopic surgery in a scalable manner, where we objectively estimate force using video input only. Those force estimations could then be used for personalized postoperative surgical quality assessment for surgeons.

To train and validate vision-based force estimation models, real-world data on retraction force applied during laparoscopic surgery on small intestines are needed but very limited [19–21]. Though some datasets using solid silicone phantoms and ad hoc animal tissue are available in the context of surgical palpation and retraction [18, 22], matching force-video dataset for the intestines is still lacking. Within the scope of this study, it must be noted that datasets captured using solid block phantoms may not provide similar mechanical responses to the intestines. Considering the easily deformable, thin-walled, and hollow-lumen structure of the small intestine, there will be a large domain gap if CV models are trained on block phantoms then directly translated to our context without thorough fine-tuning on intestinal data. Thus, effective real-world data collection and the development of data-efficient algorithms with generalization capabilities are paramount for this emerging field.

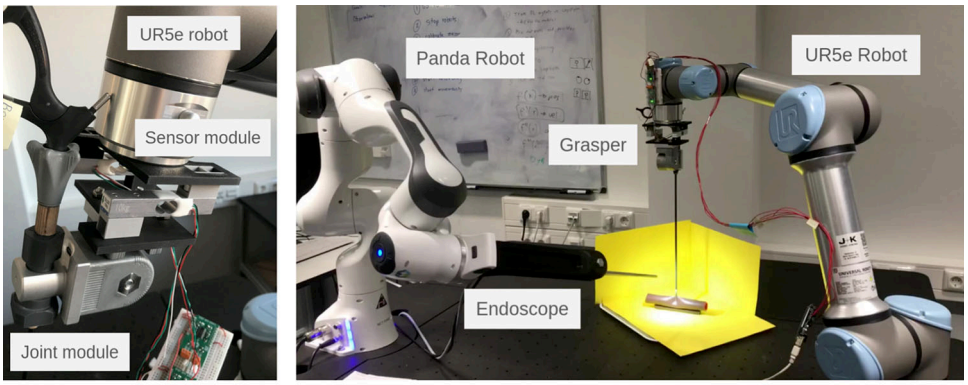
In this paper, our contributions are:

- a data collection pipeline agnostic to the surgical setting and technique, capturing videos from different camera angles with randomized pulling motions to enhance dataset variability,
- a video-force dataset with silicone small intestine phantom to enable model training—where it could be used as benchmark or pretraining data for future research,
- successful generalization capabilities of vision-based force estimation models on a preliminary dataset, evaluating their performance on unseen camera perspectives and varying bowel geometries from limited data.

All data, hardware design, and code are available upon publication here: https://gitlab.com/nct_tso_public/smallintestine-force-estimation.

Related works

Vision-based force estimation methods are less common for laparoscopy than for RAS. This is mainly due to the absence of precise instrument position tracking and robotic state data to measure force in conventional laparoscopy. Previous research has broadly divided these methods into



(a) (b)

Fig. 1 Experimental setup in detail. **a** Mechanical design of the sensor module and the grasper joint. **b** Full setup with background and a phantom

interpolation-based, learning-based methods and mixed-approach methods [23]. Interpolation-based methods capture and track surface features for geometry reconstruction [24]. These models usually involve calculations with prescribed physical models [25], pre-measured deformation-force relationships [26, 27], or finite-element analysis [28]. However, they do not take computational cost into account, despite computational hardware performance being critical for real-time clinical use. Learning-based methods do not perform any physics simulations but solely rely on a dataset with synchronized video and force signal—preferably with robot state data [29]. It has been shown that inclusion of state information increases estimation accuracy and improves generalizability when faced with unknown tools and materials [18]. Furthermore, the absence of material property input does not decrease estimation results [30]. Compared to interpolation-based methods, learning-based methods are better suited for intraoperative tasks by avoiding specifically solving the deformation problem. Generalizability is still an overall concern due to limited data realism and diversity. Mixed-approach methods amalgamate the previous categories by using CV to reconstruct deformation followed by machine learning force estimation methods. The surface reconstruction usually consists of creating point clouds and encoding temporal information with neural networks or filter blocks [17]. Unlike interpolation-based and mixed methods that rely on explicit surface reconstruction, we predict force directly from monocular video segments using vision models originally intended for other natural image applications. We compare the performance of a vision transformer and ResNet architectures, evaluating their ability to predict pulling force without requiring material properties or robotic state data.

Method

Section 2.1 describes the hardware setup including the phantom, the robots, and the sensors. Section 2.2 describes how video and force data are procured and synchronized, followed by implementation details on the CV models in Sect. 2.3. A description of experiments is presented in Sect. 2.4.

Experimental setup

To generate data for vision-based force estimation, we aim to mimic a laparoscopic setting, including the interaction between surgical instrument and phantom bowel through pulling motions. To achieve precise and controlled force measurements, we use a robotic arm for controlling the instruments, in order to accurately regulate the pulling distance and applied force.

For data capturing, a background is constructed by perpendicularly joining three large pieces of yellow fat-colored cardboard. A silicone bowel phantom (Kroton Medical Technology, Warsaw, Poland) is positioned centrally in front of the background.

The robot assembly is composed of an UR5e robot arm (Universal Robots, Odense, Denmark), a force-sensor module, and a standard laparoscopic surgical grasping instrument (Aesculap AG, Tuttlingen, Germany) secured in a joint module. We opt for a standard, non-wristed surgical instrument to collect data specifically for conventional laparoscopic surgery, as this is the most common technique for minimally invasive gastrointestinal surgery [7], but our data collection method is agnostic to the chosen surgical technique, organ, and instrument. Additionally, the method and setup can also be used for in general haptic applications, such as in [31]. The joint module involves two lockable halves and ensures stability while allowing the instrument to be angled (Fig. 1a). The joint module is attached to the force-sensor module, which is then fixed to the robot end-effector (Fig. 1b). Redesigned from the setup of [32], the force-sensor module consists of two TAL220 load cells (HT Sensor Technology, Xi'an, China) mounted on robust 3D-printed casings. The load cells are secured diagonally to ensure the measurement axis is aligned with the movements. The force signals are collected via two HX711 amplifiers and an Arduino microcontroller

(BCMI Labs SA, Chiasso, Switzerland) at 11 Hz in considerations for reduced noise.

The robot arm is controlled via a Robot Operating System (ROS) controller publishing its state at 500 Hz. A ROS topic communicates the pulling translation of the arm along with defined movement durations. Meanwhile, force data from the Arduino are published to another ROS topic. The video data are captured with a TIPCAM1 S 3D endoscope (Karl Storz SE & Co. KG, Tuttlingen, Germany) mounted on a Panda robot arm (Franka Robotics GmbH, Munich, Germany) for ease of adjustments between, and stability through experiments. Although we recognize that stereo vision is available in most surgical robotic systems and studies [17, 33], we still choose monocular vision as input to the CV model as it is more widely used in conventional laparoscopic surgery, especially considering the cost, additional setup, discomfort associated with 3D glasses, and limited rotational capabilities of angled stereo endoscopes. Hence, we have captured monocular video data to reflect the more widespread use of 2D video compared to 3D video [34]. Output images are published in a separate topic with a resolution of 960x540 pixels at 30 Hz. Since our force estimation models only require monocular video as input, only the right camera frame is used. All data are saved with time stamps.

Since the force signal is published at a different frequency than the video, all ROS topics are temporally aligned with force-signal time stamps serving as the reference. The closest video frame is selected based on the time stamps of the data—resulting in videos approximately 130 frames in length after synchronization.

Data collection and preprocessing

We start the data collection by positioning the UR5e robotic arm at a predefined initial state. At the start of each trial, the robot pivots at the grasping point by adjusting its pitch and yaw angles. The pivot is essential to ensure that each video captures pulling motions in a different direction. The pitch angle is randomly selected from a range of $-15°$ and $1.71°$, while the yaw angle is sampled within $-1.71°$ and $15°$. These specific angle constraints were determined based on the physical setup of the cardboards, where exceeding these limits could cause collisions. By restricting the pitch and yaw angles within these ranges, we maintain safe and controlled movement while maximizing the available motion space. It should also be noted that the grasper has never obscured the phantom in pulling motions, nor has any retraction taken place along the depth axis of the endoscope, leading to visual ambiguities. After this process, the load cells are zeroed to eliminate environmental force and moments. Then, the video recording starts.

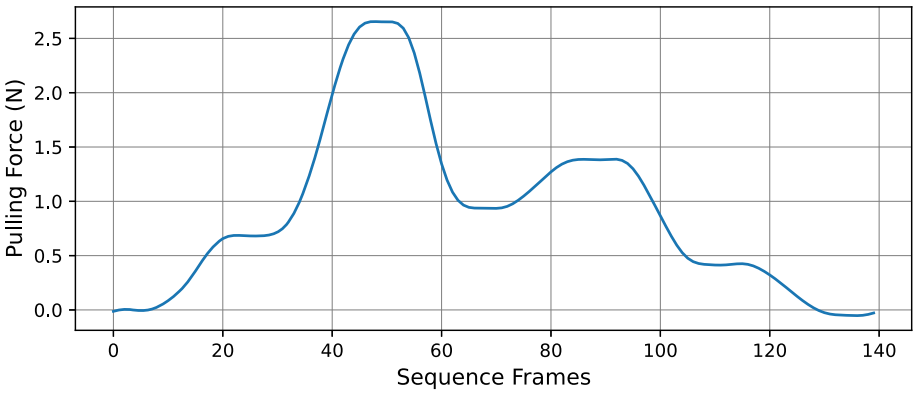
A pulling distance and a movement duration are randomly chosen within their limits. The pulling distance constraints are 3 mm and 2 cm. On the other hand, the movement duration is sampled between 1 to 2 s. Given the selected parameters, the robot then executes the pulling motion accordingly. Upon reaching the target position, the system remains stationary for a sampled duration between 0.1 and 0.5 s before proceeding to the next movement. The maximum pulling force that is recorded corresponds to approximately 5 N.

The above process is repeated five times in each recording. In each iteration, a new pulling distance and movement time combination is sampled while keeping the pitch and yaw angles fixed, introducing variability into the dataset. After completing the last pulling motion, the robot goes back to the position where no pulling force is applied. An example of a force trajectory of one recorded video can be seen in Fig. 2a.

Before the data can be passed to the network, we define the pulling motion to be of the positive force direction, inverting any force signals otherwise. The net pulling force is therefore the sum of the two load cell readings. Due to isolated large erroneous signal spiking, we cut off values at -2 and 10 N, as we observe normal recorded signal trajectory values should never be outside this range, taking into account that noisy trajectories may situate beyond 0 and 5 N. Any value beyond the cutoff lines is revised to the average of the previous and the following readings. The same correction is applied when a force reading is found to be more than 1.5 times the value of its previous counterpart—stemming from the knowledge that no defined robotic movements in the experiments could lead to such drastic accelerations.

After removing improbable sensor readings, the signals undergo fast Fourier transform and a cutoff of 0.15 cycles/frame is defined. The transformed spectrum above 0.15 or below -0.15 cycles/frame is nullified to remove high-frequency noise. At last, the inverse transformation produces the filtered signal.

In all, a dataset of 50 videos has been recorded. The videos are cropped into 540×540 pixels before resized to 112x112 pixels. A clear view of the grasper is guaranteed when manually positioning endoscope. Each video is given a random camera position, but the grasper is centrally positioned at the start of each video and its motions are in view at all times. The endoscope is always at an oblique angle to the grasper, resembling the instrument setup in a standard minimally invasive surgery with minimal visual ambiguity. It should also be noted that the camera angles ensure only the constructed background, the grasper, and the phantom are visible after data processing, as shown in Fig. 2b. Additionally, only one pulling direction is represented per video. The dataset is divided into sets of ten recordings with each set corresponding to one bowel geometry. The five bowel geometries include a straight bowel and four bent counterparts with randomized angles between within 30° (6°, 16°,



(a)

(b)

Fig. 2 Force and video collections. **a** Pulling force trajectory of a recorded video. **b** Exemplary frame of a recorded video. The red square shows the applied crop

21°, 25°), maintaining a circular lumen along the length of the phantom.

Network architectures and training

Three vision-based force estimation architectures are trained: 3D-ResNet-18, (2+1)D ResNet [35], and Video Swin Transformer [36], with 3D-ResNet-18 serving as the baseline model. A dataloader based on the design of [31] reads video data and force ground truth. A major advancement, thanks to the models to inherently process video data, is that once the data are loaded, a temporal window of defined length is read as a unit and rolls to later frames, eliminating the necessity of additional modules such as LSTM or RNN to capture temporal information [18].

We change the final fully connected layer of the models to produce a regression output instead of a classification output. The feature vector obtained from the vision models is first passed through a fully connected layer that reduces its dimensionality to 512 neurons, followed by the application of a ReLU activation function to introduce nonlinearity. Subsequently, a dropout rate of 0.5 is employed to prevent overfitting. Finally, the processed neurons are input into a fully connected layer that outputs a one-dimensional vector corresponding to the predicted force. Training the three models using mean-square-error (MSE) loss and a batch size of 5 for 50 epochs shows that the best learning rate is 10^{-4}. We train the models using mixed precision on one Nvidia RTX A5000 GPU.

For the three models, we initialize training using pretrained weights from the Kinetics-400 dataset [37]. In the interest of memory requirements and generalizability, we freeze the initial layers during training. In the case of the 3D-ResNet-18 and (2 + 1)D, we fine-tune the third and fourth blocks and the regression head. For the transformer, we fine-tune the fifth and sixth feature blocks, the final normalization layer, and the regression head. To mitigate the risk of overfitting, we apply L2 regularization to the trainable weights of the model, using a weight decay equal to 10^{-2}. The data loading and training regime described above is also illustrated in Fig. 3.

To encourage gradual convergence, we use cosine annealing, with a minimum value of 10^{-5} as the learning rate decay schedule with the AdamW optimizer, providing a stable fine-tuning process. During training, data augmentation, such as horizontal flip, Gaussian blur, and RGB shift, was applied to

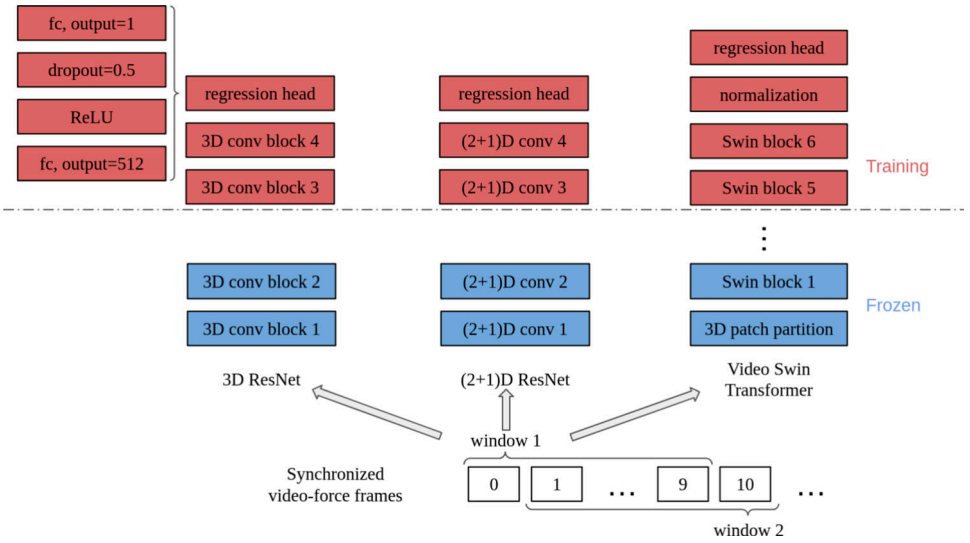


Fig. 3 Data loading and training: the rolling window method is illustrated with assumption that window length is set to 10 frames. Each loaded window is then passed into a model for training, where the first blocks are frozen and only the upper feature blocks and regressions heads are trained

the input sequences, as we find the models exhibit tendencies to overfit in its absence, reducing generalizability.

Experimental design

We design two fivefold cross-validation experiments, selecting root-mean-square error (RMSE) as evaluation metric. Additionally, normalized RMSE (NRMSE) is also reported with respect to the total range of applied forces in the test sets, as calculated in equation 1. Primary assessments in the generalizability of the models are undertaken in two key aspects: unseen phantom geometries and unseen camera angles. In both experiments, we keep a ratio of 60% of the data for training, 20% for validation, and 20% for testing.

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(F_{\text{estimated},\,t} - F_{\text{measured},\,t}\right)^2}}{F_{\max} - F_{\min}} \tag{1}$$

The first experiment addresses the challenge of unseen phantom geometries. In each fold of this experiment, one bowel geometry is completely excluded from training and used solely for testing. From the remaining four geometries, two videos per geometry are randomly selected for validation, while the remaining videos are used for training.

For the second experiment evaluating generalization to unseen camera angles, the data are split differently. As every video in the dataset has a unique camera angle, we select two videos from each of the five geometries for validation and two for testing, while the remaining six videos are used for training. In the first fold, videos 1 and 2 from each geometry are assigned to the test set, while videos 3 and 4 are used for validation. In the second fold, videos 3 and 4 are used for testing, while videos 5 and 6 are assigned to validation. This process continues in subsequent folds, ensuring that every video is eventually included in both validation and test sets across different iterations.

In the scope of the first and second experiments, we also report the inference time of every model with different window lengths. With this measurement, we show the inference speed capabilities of the methods.

To compare our method with existing CV models, we additionally conducted two benchmark experiments. In the third experiment, a dataset published by Chua et al, containing both retractions and palpations [18], is selected. In order to follow the evaluation techniques of the benchmark method, modifications are made where the last fully connected layer of each model has been replaced to output three values representing forces in $x, y, andz$ directions. To maintain consistency with the benchmark, center cropping was performed at 300×300 pixels as specified in [18]. The cropped images were then resized to 112×112 pixels to match the input dimensions of our models. All training hyperparameters were maintained consistent with those used for training on our dataset. The models were trained for 50 epochs with the same dataset divisions outlined in [18] subject to dataset completeness.

The fourth experiment, using the same model modifications specified in the last, is also found upon it. A more realistic dataset has been used, which contains manipulation recordings with raw chicken tissue instead of silicone setups [22]. In the benchmark experiment, images are cropped to windows of 234×234 pixels, centered on the M2 keypoint. However, since our pipeline does not incorporate keypoint detection, we retain the 300×300 center crop. The model from the third experiment has been further fine-tuned on the realistic dataset for 50 epochs maintaining the exact same hyperparameters. The results from this experiment will be able to indicate the models' generalizability to data from a different domain.

Table 1 Test RMSE (N) and NRMSE for geometry experiments

Metric	Window length	3D ResNet	(2 + 1)D ResNet	Video Swin Transformer
RMSE (N)	10	**0.323 ± 0.066**	0.327 ± 0.068	0.574 ± 0.087
	20	0.342 ± 0.074	**0.339 ± 0.059**	0.469 ± 0.065
	30	0.374 ± 0.100	**0.340 ± 0.061**	0.468 ± 0.081
NRMSE	10	**0.077 ± 0.016**	0.078 ± 0.017	0.136 ± 0.023
	20	0.081 ± 0.017	**0.080 ± 0.014**	0.111 ± 0.015
	30	0.088 ± 0.019	**0.081 ± 0.014**	0.110 ± 0.014

Best results for each window length are highlighted in bold

Table 2 Test RMSE (N) and NRMSE for camera angle experiments

Metric	Window length	3D ResNet	(2 + 1)D ResNet	Video Swin Transformer
RMSE (N)	10	**0.318 ± 0.050**	0.331 ± 0.102	0.518 ± 0.059
	20	**0.313 ± 0.028**	0.332 ± 0.071	0.425 ± 0.050
	30	0.328 ± 0.057	**0.321 ± 0.033**	0.404 ± 0.064
NRMSE	10	**0.070 ± 0.011**	0.075 ± 0.025	0.115 ± 0.015
	20	**0.069 ± 0.007**	0.073 ± 0.015	0.094 ± 0.012
	30	0.072 ± 0.013	**0.071 ± 0.004**	0.089 ± 0.015

Best results for each window length are highlighted in bold

Evaluation

Throughout the first and second experiments, test results from Tables 1 and 2 show that the ResNet-based models are able to estimate applied force significantly more accurately than the Video Swin Transformer (paired t test, $p < 0.0001$). However, the same test ($p = 0.90$) also shows that (2+1)D ResNet demonstrates no superior performance than 3D ResNet, nor vice versa. We speculate that these findings are due to transformers being associated with worse performance than convolutional neural networks when training data are scarce, as also seen in the original Vision Transformer [38]. It is important to note at this point that training the full transformer model might improve its performance, with the trade-off of memory consumption and training-time. The trajectories in Fig. 4 illustrate this point where the transformer trajectory could follow the dynamics of the ground truth, but is often separated from the ground truth by a margin. Noting that standard deviations in Table 1 across experiments are generally only around 20%, we conclude that the stable results from cross-validation experiments indicate that all our models have performed well in the perspective of generalization over phantom geometry and camera angles. Thus, the main assessments of the experiments have been satisfied.

We ran further tests where we show that all three models are real-time capable, as Table 3 contains models' inference time per output frame. In order to provide real-time visual force feedback, such as an overlay on the surgical video, the model update rate should match the video frame rate (larger than 25 frames per second) with system delay under 100ms for unhindered operations [39]. The results demonstrate that 3D ResNet stands as the fastest model, while Video Swin Transformer is the slowest. It is also confirmed that, independently of the model, with smaller window length the inference time is faster which it is an expected result. That said, the slowest condition of all scenarios offered 0.034 s per frame, corresponding to approximately 29Hz, which is above the data collection rate. Therefore, it can be concluded that all models can perform real-time inference with our setup.

This finding indicates that our method enables the development of systems for building real-time force feedback systems. In this case, additional hardware systems providing haptic feedback similar to the Da Vinci 5 system are further required. However, the setup cannot currently be implemented clinically without regulatory obstacles. We see, however, potential in providing visual force feedback in lieu, which can be deployed on a large scale.

While investigating whether varying input window lengths affect testing results using paired t tests, we find that using 10-frame windows leads to worse results than cases with 20-frame windows ($p = 0.02$). However, models using 30-frame windows produce insignificant improvements ($p = 0.73$) than results with 20-frame window inputs. Hence, we conclude that 20 frames is the ideal input sequence window length where enough temporal information has been included while not requiring additional memory.

To compare our methods to others in the third and fourth experiments, we train our models with a window length of 20. Table 4 demonstrates results from the third experiment using silicone manipulation data. Since our model has only visual inputs due to considerations on clinical implementation, the benchmark RMSE values are chosen from the vision-input

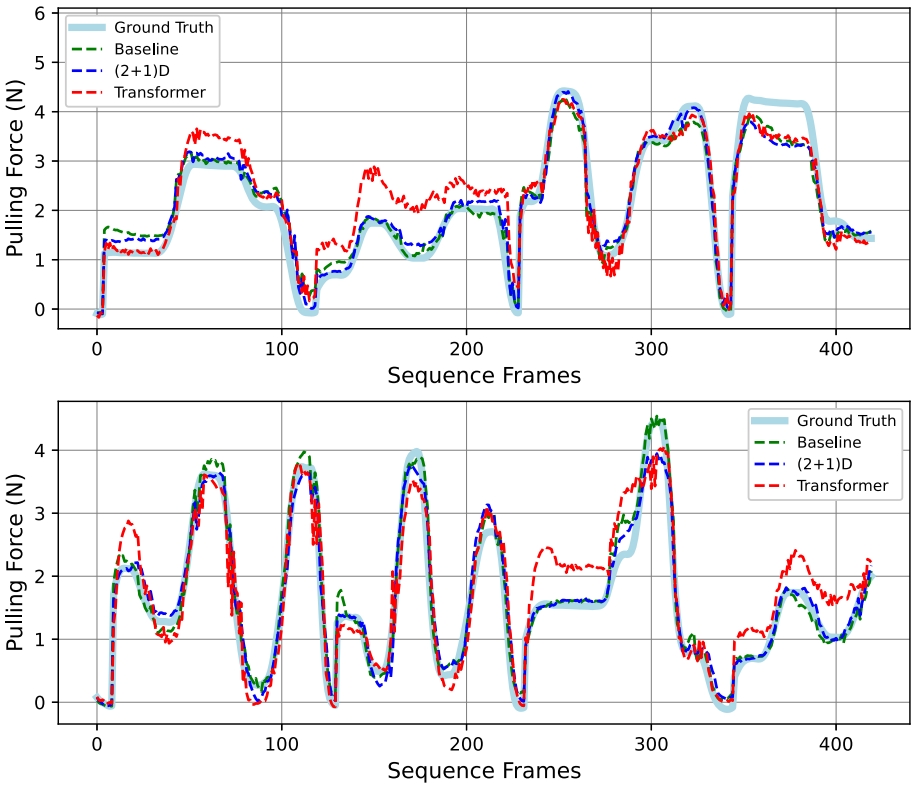


Fig. 4 Two segments of predicted force trajectories from testing dataset with 3D and (2 + 1)D ResNets having higher accuracy compared to the transformer model

Table 3 Inference time (secs) against window lengths

Model	10 frames	20 frames	30 frames
3D ResNet	**0.011**	**0.018**	**0.023**
(2 + 1)D ResNet	0.011	0.021	0.028
Video Swin Transformer	0.023	0.030	0.034

The bold values refer to best performance in each column

only models in the work of Chua et al. [18]. It is also shown in Table 4 that the ResNet-based networks provide for higher force estimation accuracy than the Video Swin Transformer and the referenced benchmark method in workspace and viewpoint configurations, as well as the unseen scenarios. Given such performance, it can be concluded that the ResNet-based methods should consistently perform better than the transformer and the benchmark method. Comparing UM (Unseen Material), UT (Unseen Tool), and mean configuration results from Table 4 further show that the models perform well to seen workspace and viewpoint configurations, but not to the two unseen cases, where they could lead to double the error. Therefore, complete retraining on a new dataset is advisable to achieve the best result when the test setup or material is different from that of training. Moreover, in considerations of generalization to increase accuracy in unseen environments, we suggest adding diversity to the data in terms of background, retraction/palpation, and surgical instrument.

To examine how our models perform with complex background, Table 5 demonstrates results from the fourth experiment with the realistic dataset using chicken tissue. We benchmark our models with the graph neural network (GNN) and fully connected neural network (FCN) trained by [22]. It is shown that our models outperform the referenced GNN and FCN methods in the silicone dataset and in the Y-axis of the realistic dataset, but not in the X- and Z-axis—nevertheless—the differences appear small. We infer that this result is due to the knowledge of the past that our models retain from the window input approach, which the benchmarks lack. We notice that model performance degrades significantly when using the realistic dataset. It indicates that data complexity is a factor of estimation accuracy, and it may not be wholly compensated by retraining or fine-tuning. We speculate that adding other input modalities, such as robot state, might improve the results [18].

Conclusion

This work contributes to advancing scalable surgical quality assessment by demonstrating that computer vision can serve as a reliable surrogate for direct force sensing in tissue manipulation. We introduce the use of video as a primary modality for objectively estimating tissue retraction force, a novel approach that addresses the lack of haptic feedback in

Table 4 RMSE (N) across workspace and viewpoint configurations

Model	Workspace and viewpoint configuration											UM	UT
	L3	L2	L1	C	R1	R2	R3	Z1	Z2	Z3	Mean		
3D ResNet	0.421	**0.362**	0.362	0.435	0.318	0.347	0.418	**0.552**	0.455	0.501	0.417	**1.101**	0.774
(2 + 1)D ResNet	**0.407**	0.346	**0.334**	**0.417**	**0.293**	**0.340**	**0.404**	0.556	**0.448**	**0.491**	**0.404**	1.151	**0.727**
Video Swin Transformer	0.467	0.437	0.443	0.535	0.382	0.408	0.492	0.634	0.563	0.651	0.501	1.177	0.799
Chua et al	0.521	0.499	0.504	0.496	0.456	0.499	0.465	0.721	0.728	0.714	0.560	1.837	0.740

The bold values refer to best performance in each column

Table 5 Test NRMSE for fine-tuning on silicone and realistic dataset

Model	Silicone			Realistic		
	X	Y	Z	X	Y	Z
3D ResNet	0.062 ± 0.022	0.084 ± 0.025	0.053 ± 0.033	0.102 ± 0.025	**0.114 ± 0.034**	0.106 ± 0.033
(2 + 1)D ResNet	**0.061 ± 0.026**	**0.082 ± 0.024**	**0.050 ± 0.032**	0.101 ± 0.025	0.115 ± 0.030	0.106 ± 0.032
Video Swin Transformer	0.078 ± 0.024	0.091 ± 0.024	0.060 ± 0.031	0.105 ± 0.023	0.121 ± 0.034	0.108 ± 0.026
GNN [22]	0.115 ± 0.014	0.126 ± 0.015	0.093 ± 0.022	**0.085 ± 0.014**	0.130 ± 0.025	0.108 ± 0.030
FCN [22]	0.116 ± 0.014	0.121 ± 0.015	0.093 ± 0.026	**0.085 ± 0.015**	0.127 ± 0.021	**0.103 ± 0.024**

The bold values refer to best performance in each column

minimally invasive surgery. A novel low-cost robotic setup has been presented to collect retraction data for training CV models. Additionally, a data collection pipeline, agnostic to surgical setting and operating technique, led to a dataset with videos from different camera angles with randomized pulling motions. We then demonstrated successful generalization capabilities of the vision-based force estimation models on the dataset, showing small estimation errors and real-time inference capabilities.

Comparing the RMSE values with previous force estimation works, we find that our errors are comparable to or lower than those reported in the literature. However, direct comparison in the case of small bowel is unavailable—and for this reason we will release our dataset. Results show that our models are able to estimate pulling force with the phantom—which is more fragile and deformable than solid silicone phantoms previously used [30]. It should be further noted that our models use vision alone, without referring to depth estimation [40], robotic state [18], or physical simulations [28]. This design feature reflects our considerations in that tracking non-robotic laparoscopic tools is very technically demanding and measured positions are usually not precise enough to support force estimation. Hence, out methods remove the dependency on positional information, making them suited for clinical adoption without the necessity of modifying existing systems. However, the lack of the other input sources could limit generalizability with reduced estimation accuracy when visually similar yet different anatomies are concerned. Furthermore, the methodological decision to use monocular vision instead of stereo vision makes scene understanding more challenging, as it lacks access to absolute depth maps and surface reconstructions—the effect of all of which would be meaningful future studies.

While models presented in this work show promising results, further investigation could proceed by adding additional model knowledge with the inclusion of a local depth-vision keypoint tracking module. Furthermore, the motion in the dataset is restricted to retraction, which could confine the model to this task. It is therefore meaningful to continue exploring how networks will perform in *ex vivo* and *in vivo* environments by collecting more surgical datasets [30]. It should be noted that any clinical feedback implementation will only be possible after further validation with *in vivo* data, following which could a response platform be built to provide assessments in training or to enable direct intraoperative haptic feedback. Furthermore, though recorded inference times imply real-time model capabilities, it is important to note that actual deployment is affected by delays in video frame acquisition, buffering, and overhead times of the systems that make such performance an upper bound rather than a guarantee.

Acknowledgements This work is supported by funding from the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of TUD Dresden University of Technology. The authors would like to thank the Federal Ministry of Research, Technology, and Space (BMFTR) for its support as part of the research program Communication Systems "Souverän. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KIS2413K.

Author Contributions K. Wang implemented data preprocessing and the force sensor module. A. Rodriguez designed the robot control algo-

rithm and data synchronization. Experimental design and data collection were shared equally between the aforementioned authors. R. Younis revised the work critically. M. Pfeiffer, S. Bodenstedt, M. Wagner, and S. Speidel supervised and revised this work.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors identify no conflict of interest.

References

1. Cameron JL (1997) William Stewart Halsted: our surgical heritage. Ann Surg 225(5):445–458
2. Tang B, Hanna G, Joice P, Cuschieri A (2004) Identification and categorization of technical errors by observational clinical human reliability assessment (OCHRA) during laparoscopic cholecystectomy. Arch Surg 139(11):1215–1220
3. Orr KE, Williams MP (2014) MDCT of retractor-related hepatic injury following laparoscopic surgery: appearances, incidence, and follow-up. Clin Radiol 69(6):606–610. https://doi.org/10.1016/j.crad.2014.01.008. (**Accessed 2025-02-08**)
4. Roca E, Ramorino G (2023) Brain retraction injury: systematic literature review. Neurosurg Rev 46(1):257. https://doi.org/10.1007/s10143-023-02160-8. (**Accessed 2025-02-08**)
5. Leevan E, Carmichael JC (2019) Iatrogenic bowel injury (early vs delayed). In: Seminars in colon and rectal surgery, vol 30. Elsevier, p 100688
6. Mendez LE (2001) Iatrogenic injuries in gynecologic cancer surgery. Surg Clin North Am 81(4):897–923
7. Sheetz KH, Claflin J, Dimick JB (2020) Trends in the adoption of robotic surgery for common surgical procedures. JAMA Netw Open 3(1):1918911–1918911
8. Wottawa CR, Genovese B, Nowroozi BN, Hart SD, Bisley JW, Grundfest WS, Dutson EP (2016) Evaluating tactile feedback in robotic surgery for potential clinical application using an animal model. Surg Endosc 30(8):3198–3209
9. Muscolo GG, Fiorini P (2023) Force-torque sensors for minimally invasive surgery robotic tools: an overview. IEEE Trans Med Robot Bionics 5(3):458–471
10. Hao Y, Zhang H, Zhang Z, Hu C, Shi C (2024) Development of force sensing techniques for robot-assisted laparoscopic surgery: a review. IEEE Trans Med Robot Bionics 6(3):868– 887. https://doi.org/10.1109/TMRB.2024.3407238
11. Hagen M, Meehan J, Inan I, Morel P (2008) Visual clues act as a substitute for haptic feedback in robotic surgery. Surg Endosc 22:1505–1508
12. Golahmadi AK, Khan DZ, Mylonas GP, Marcus HJ (2021) Tool-tissue forces in surgery: a systematic review. Ann Med Surg 65:102268
13. Sugiyama T, Lama S, Gan LS, Maddahi Y, Zareinia K, Sutherland GR (2018) Forces of tool-tissue interaction to assess surgical skill level. JAMA Surg 153(3):234–242
14. Bergholz M, Ferle M, Weber BM (2023) The benefits of haptic feedback in robot assisted surgery and their moderators: a meta-analysis. Sci Rep 13(1):19215
15. Martin J, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273–278
16. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113
17. Aviles AI, Alsaleh SM, Hahn JK, Casals A (2016) Towards retrieving force feedback in robotic-assisted surgery: a supervised neuro-recurrent-vision approach. IEEE Trans Hapt 10(3):431–443
18. Chua Z, Jarc AM, Okamura AM (2021) Toward force estimation in robot-assisted surgery using deep learning with vision and robot state. In: 2021 IEEE international conference on robotics and automation (ICRA). IEEE, pp 12335–12341
19. Miyasaka EA, Okawada M, Utter B, Mustafa-Maria H, Luntz J, Brei D, Teitelbaum DH (2010) Application of distractive forces to the small intestine: defining safe limits. J Surg Res 163(2):169–175
20. Shah D, Alderson A, Corden J, Satyadas T, Augustine T (2018) In vivo measurement of surface pressures and retraction distances applied on abdominal organs during surgery. Surg Innovat 25(1):50–56
21. Durcan C, Hossain M, Chagnon G, Perić D, Girard E (2024) Mechanical experimentation of the gastrointestinal tract: a systematic review. Biomech Model Mechanobiol 23(1):23–59
22. Yang S, Le MH, Golobish KR, Beaver JC, Chua Z (2024) Vision-based force estimation for minimally invasive telesurgery through contact detection and local stiffness models. J Med Robot Res 9(03n04):2440008
23. Nazari AA, Janabi-Sharifi F, Zareinia K (2021) Image-based force estimation in medical applications: a review. IEEE Sens J 21(7):8805–8830
24. Wang J, Yao M, Wei Y, Guo X, Zheng A, Zhao W (2025) Image-to-force estimation for soft tissue interaction in robotic-assisted surgery using structured light. IEEE Robot Autom Lett 10(8):7795–7802. https://doi.org/10.1109/LRA.2025.3579640
25. Haouchine N, Kuang W, Cotin S, Yip M (2018) Vision-based force feedback estimation for robot-assisted surgery using instrument-constrained biomechanical three-dimensional maps. IEEE Robot Autom Lett 3(3):2160–2165
26. Lin W-C, Song K-T (2016) Instrument contact force estimation using endoscopic image sequence and 3d reconstruction model. In: 2016 international conference on advanced robotics and intelligent systems (ARIS). IEEE, pp 1–6
27. Noohi E, Parastegari S, Žefran M (2014) Using monocular images to estimate interaction forces during minimally invasive surgery. In: 2014 IEEE/RSJ international conference on intelligent robots and systems. IEEE, pp 4297–4302
28. Xie H, Song J, Zhong Y, Gu C, Choi K-S (2022) Constrained finite element method for runtime modeling of soft tissue deformation. Appl Math Model 109:599–612
29. Jung W-J, Kwak K-S, Lim S-C (2020) Vision-based suture tensile force estimation in robotic surgery. Sensors 21(1):110
30. Lee Y-E, Husin HM, Forte M-P, Lee S-W, Kuchenbecker KJ (2023) Learning to estimate palpation forces in robotic surgery from visualinertial data. IEEE Trans Med Robot Bionics 5(3):496–506. https://doi.org/10.1109/TMRB.2023.3295008

31. Shin H, Cho H, Kim D, Ko D-K, Lim S-C, Hwang W (2019) Sequential image-based attention network for inferring force estimation without haptic sensor. IEEE Access 7:150237–150246
32. Petrea RAB, Bertoni M, Oboe R (2021) On the interaction force sensing accuracy of Franka Emika panda robot. In: IECON 2021—47th annual conference of the IEEE Industrial Electronics Society. IEEE, pp 1–6
33. Aviles AI, Marban A, Sobrevilla P, Fernandez J, Casals A (2014) A recurrent neural network approach for 3d vision-based force estimation. In: 2014 4th international conference on image processing theory, tools and applications (IPTA), pp 1–6. https://doi.org/10.1109/IPTA.2014.7001941
34. Sørensen SMD, Savran MM, Konge L, Bjerrum F (2016) Three-dimensional versus two-dimensional vision in laparoscopy: a systematic review. Surg Endosc 30(1):11–23
35. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6450–6459
36. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022) Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3202–3211
37. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev A, Suleyman M, Zisserman A (2017) The kinetics human action video dataset. arXiv:abs/1705.06950
38. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations. https://openreview.net/forum?id=YicbFdNTTy
39. Xu S, Perez M, Yang K, Perrenot C, Felblinger J, Hubert J (2014) Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dv-trainer® simulator. Surg Endosc 28(9):2569-2576
40. Gao C, Liu X, Peven M, Unberath M, Reiter A (2018) Learning to see forces: surgical force prediction with rgb-point cloud temporal convolutional networks. In: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis: first international workshop, OR 2.0 2018, 5th international workshop, CARE 2018, 7th international workshop, CLIP 2018, 3rd international workshop, ISIC 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5. Springer, pp 118–127