Bayesian Temporal Pose Networks for Uncertainty-Calibrated Laparoscopic Tool Pose Tracking

Omar Choudhry[1], Sharib Ali[1], Chandra Shekhar Biyani[2], and Dominic Jones[3(✉)]

[1] School of Computer Science, University of Leeds, Leeds, UK
[2] St James's University Hospital, Leeds Teaching Hospitals NHS Trust, Leeds, UK
[3] School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK
{O.Choudhry, S.S.Ali, D.P.Jones}@leeds.ac.uk, shekhar.biyani@nhs.net

Abstract. Laparoscopic instrument pose tracking from monocular endoscopic video in surgical training tasks is essential for computer-assisted surgery and objective skill assessment. However, current methods require geometric priors unavailable in non-robotic settings and lack temporal reasoning across multimodal cues and uncertainty quantification. We introduce Bayesian Temporal Pose Network (BTPN), a framework that fuses visual and kinematic features through hierarchical multi-scale temporal attention with calibrated Bayesian uncertainty. A fine-tuned segmentation backbone achieves 99.1 % mAP_{50} and keypoint detection reaches 98.3 % mAP_{50}. End-to-end visual pose tracking attains 7.0 mm position and 11.7° rotation RMSE with 0.028 uncertainty error. Our code is available at https://github.com/omariosc/BTPN.

Keywords: computer vision · image processing · learning (artificial intelligence) · uncertainty · tool tracking · pose estimation · surgery

1 Introduction

Real-time pose tracking of laparoscopic instruments from monocular endoscopic video of surgical training tasks is a challenging yet crucial task in computer-assisted surgery. Accurate 6-DoF (Degrees of Freedom) pose (position and orientation) can enable augmented reality (AR) guidance, safety monitoring, and objective skill assessment in minimally invasive surgery [6, 14]. However, recovering full 6D pose remains an open problem with prior approaches relying on geometric priors unavailable in non-robotic settings [1]. Computer-Aided Design (CAD)-dependent models achieve state-of-the-art 6D pose accuracy but require a 3D model of each target object: FoundationPose [21] uses either a CAD mesh or reference images, while some recent work obtains geometry via training on synthetic renderings [20]. Robot-specific methods exploit forward kinematics and known link geometry; the SurgRIPE benchmark [23] is the first robotic dataset to provide joint 3D pose and segmentation masks. Geometric assumptions relax

the CAD requirement but impose shape constraints [19]: ART-Net [10] recovers 3D pose via algebraic geometry under a cylindrical shaft assumption. Where priors are absent, existing work is confined to 2D outputs where YOLO models have been used [4, 12]. The PhaKIR 2024 keypoint estimation challenge [15] only received two entries, both degrading severely under occlusion. Non-robotic laparoscopic tools lack CAD models, kinematic chains, and standardised geometry. The clinically relevant point of interest is the jaw tooltip rather than the shaft, precluding direct application of any existing method. Existing approaches also largely operate using markers [9] or naïve bounding-box tracking [1, 3] on individual frames, lacking other innovations in surgical data science, e.g. temporal reasoning [2, 25] or uncertainty quantification [7, 17], essential for clinical trust [13, 16].

Our work addresses this gap by presenting the Bayesian Temporal Pose Network (BTPN), a framework for probabilistic 7-DoF tool pose tracking from monocular video, requiring no geometric or CAD priors. Our primary contributions are: **(i)** A hierarchical multi-scale temporal attention mechanism operating at clinically motivated resolutions with quaternion-aware bimanual cross-attention for capturing inter-tool coordination and learnable cross-scale fusion using segmentation (91.1 % $\text{mAP}_{50\text{-}95}$) and keypoint estimation (94.6 % $\text{mAP}_{50\text{-}95}$); **(ii)** A kinematic foundation model that learns tool dynamics by predicting future poses, using ∼330 K electromagnetic (EM) tracking data frames across 114 peg transfer trials.

2 Methodology

2.1 Datasets

We use the multimodal LASK datasets [5] of laparoscopic peg transfer trials recorded using cameras for image frames and an NDI Aurora EM tracking system for ground truth tool poses (Table 1). Approval for the dataset collection was granted by the University of Leeds Faculty of Engineering and Physical Sciences Research Ethics Committee (ref: EPS 2211 and MEEC 22-023). EM coils near each tool shaft base, calibrated to the tool tips, provide per-frame position and orientation quaternion, recorded in the Aurora tracker frame and transformed into the camera frame using the camera pose and its intrinsic and extrinsic parameters. The Aurora system has 0.48 mm position and 0.3° orientation RMSE; per-tool coil-to-tip calibration recovers the sensor offset by axial-shaft rotation, the jaw-hinge offset by a pointed-cap pivot, and axial roll against a world reference. All trials used a non-ferrous box trainer, keeping interference below the system's drift threshold. Magnetic sensors and magnets attached to the lower and upper tool handles yield voltages proportional to jaw aperture. This voltage is smoothed with a second-order Butterworth low-pass filter (0.5 Hz cutoff, zero-phase using offline forward-backwards filtering) and calibrated per trial to a $[0, 1]$ opening fraction by 10th/90th-percentile scaling; jaw error is reported as a percentage of this opening range. The datasets were collected over three years; jaw sensing was added after Dataset A (so Dataset C has none).

Table 1: Dataset overview. Each trial corresponds to one unique participant (urology surgeons for A and C, paediatric surgeons for B).

Dataset	Collected	Trials	DoF	Frames	fps	Jaw	Ann.	Neg.	Pos.	Role
A	Oct 2024	60	7	∼86 K	13	✓	4278	3836	442	Train / val
B	Nov 2024	30	7	∼48 K	13	✓	4784	4607	177	In-dist. ext. val
C	Oct 2023	24	6	∼107 K	26	—	416	108	308	OOD test

(a) Dataset A (b) Dataset B (c) Dataset C

Fig. 1: Representative annotated frames from each dataset.

Dataset A comprises 60 urology specialists; Dataset B, 30 paediatric specialists; Dataset C, 24 urology specialists collected one year earlier. One-way analysis of variance (ANOVA) confirms significant distributional differences across datasets ($p<0.001$) due to equipment differences in training boxes, camera/tool models and insertion angles, with Dataset B suitable for in-distribution validation and C for OOD (out-of-distribution). Each annotated frame contains per-tool polygon segmentation masks and 8 keypoints: proximal and distal shaft edge endpoints, end-effector joints and tips, and both jaws. Negative frames in which neither tool is present are explicitly included to reduce false-positive detections (Table 1).

2.2 Problem Formulation

Given an endoscopic video $\{I_t\}_{t=1}^{T}$, we estimate 3D poses of each laparoscopic tool $k \in \{L, R\}$ (left, right) with a calibrated measure of prediction confidence, using only the images as input at inference time. Our objective is to maximise the likelihood of the future pose $\mathbf{y}_{t+1}^{(k)}$ under the predicted distribution $p(\mathbf{y}_{t+1}^{(k)} \mid I_{1:t})$, while ensuring that the predicted uncertainty is well-calibrated with respect to observed errors, by minimising a composite loss (Sec. 2.4). The per-tool pose is $\mathbf{y}_t^{(k)} = [\mathbf{p}, \mathbf{q}, \theta]$, $\mathbf{p} \in \mathbb{R}^3$ is the 3D tool-tip position in the camera coordinate frame (mm), $\mathbf{q} \in \mathbb{S}^3$ is the orientation as a unit quaternion, and $\theta \in [\theta_{\min}, \theta_{\max}] \subset \mathbb{R}$ is the jaw angle. These components inhabit geometrically distinct spaces and require distribution families that respect each geometry.

$$p\Big(\mathbf{y}_{t+1}^{(k)} \mid I_{1:t}\Big) = \underbrace{\mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)}_{\text{position}} \cdot \underbrace{\mathrm{vMF}(\boldsymbol{\mu}_q, \kappa)}_{\text{orientation}} \cdot \underbrace{\mathcal{N}(\mu_\theta, \sigma_\theta^2)}_{\text{jaw angle}}, \tag{1}$$

By conditioning causally on all frames up to time t at multiple temporal scales (enabling it to exploit motion continuity and anticipate short-horizon tool trajectories), the network predicts a factored distribution over future frames poses,

where the position mean $\boldsymbol{\mu}_p \in \mathbb{R}^3$ and covariance $\boldsymbol{\Sigma}_p = \mathbf{L}\mathbf{L}^\top \succ 0$ are parameterised via a lower-triangular Cholesky factor $\mathbf{L}$; avoiding costly eigenvalue clamping or projection during training. For orientation, a standard Gaussian on $\mathbb{R}^4$ would assign probabilities on invalid (non-unit) quaternions; we instead adopt the von Mises–Fisher distribution $\mathrm{vMF}(\boldsymbol{\mu}_q, \kappa)$, defined natively on the unit hypersphere $\mathbb{S}^3$ and concentration parameter $\kappa > 0$ where high κ concentrates density around the mean direction $\boldsymbol{\mu}_q$ (confidence), while low κ spreads density over the sphere (uncertainty). For the jaw angle, μ_θ and σ_θ^2 parameterise a scalar Gaussian matching its 1D Euclidean nature.

2.3 Network Architecture

The three panels of Fig. 2 work together as follows. **(a) BTPN:** A two-stage pipeline avoids tool-identity confusion: YOLOv26 [11] localises and segments both tools, then per-tool ROI crops pass to a YOLOv26m-pose model for 8-keypoint detection, with a frozen DepthAnything V2 encoder [24] adding monocular depth cues along the optical axis. A matching image-based encoder produces a visual embedding (V), which a gated cross-modal fusion combines with the kinematic embedding (K) supplied by the KFM. Separate probabilistic heads then predict the position, rotation, and jaw distributions as gated residual corrections to the kinematic prior, $\hat{\mathbf{p}} = \boldsymbol{\mu}_p^{(\mathrm{kin})} + g_p\, \delta\boldsymbol{\mu}_p$. Separate per-channel confidence gates prevent noisy visual estimates from corrupting modalities where

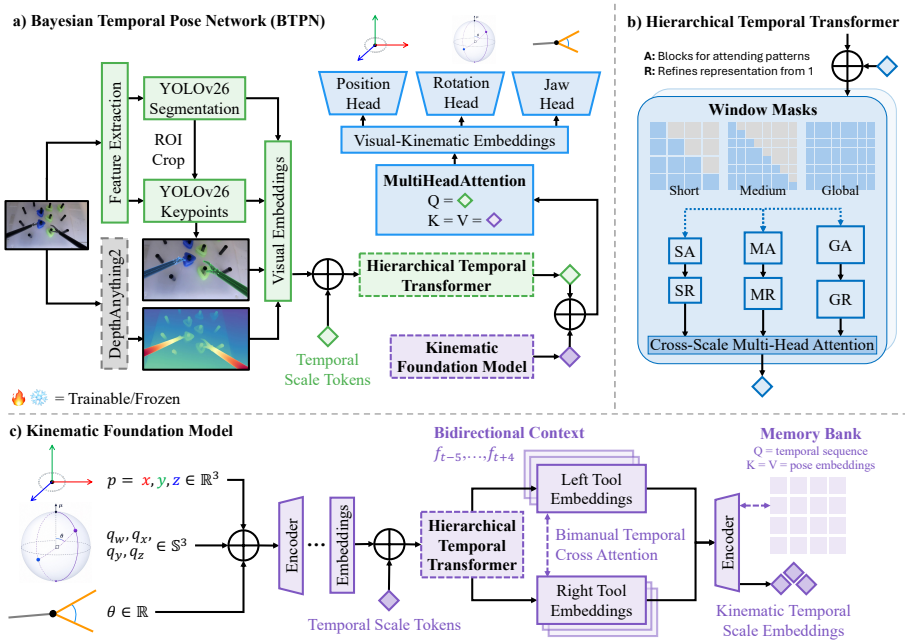

Fig. 2: BTPN architecture overview.

kinematics are already accurate (position and jaw $\leq$0.5, rotation $\leq$0.1). **(b) Hierarchical Temporal Transformer (HTT):** Used by both the visual and kinematic branches, the HTT models vision/motion through windowed attention at three clinically-motivated temporal scales spanning 10/50/100 frames ($\approx$0.8/3.8/7.7 s at 13 fps): a local scale for micro-gestures (grasp onset, release timing), a medium scale for purposeful actions, and a global scale for subtask context. These follow measured peg-transfer action durations (reach $\approx$1.1, grasp $\approx$1.2, transfer $\approx$2.4, place $\approx$2.0 s), with local micro-gestures nested within the longer phases, so each scale matches the temporal extent of a clinically meaningful motion. **(c) Kinematic Foundation Model (KFM):** An HTT trained on EM trajectories to predict the next-frame pose from a causal context window, learning tool dynamics that serve as a strong motion prior. A Memory-Enhanced Encoder (MEE) augments these embeddings with a learned memory bank — a small set of trainable prototype vectors that the sequence attends to through a gated residual, recalling recurring motion patterns beyond the current window using bidirectional context. Since surgeons do not move the two tools independently, the MEE shares the per-tool streams with self-attention in addition to quaternion-aware bimanual cross-attention: each tool first attends over its own sequence and then into the other tool's features, with a learned gate controlling how much cross-tool information flows.

2.4 Loss Functions

Eq 2: Beta-NLL (position and jaw angle) Beta-NLL (negative log-likelihood) re-weights gradients to prevent variance collapse on easy samples [18]. Here $\sigma_{\text{det}} = \text{stopgrad}(\sigma)$ (and likewise κ_{det} for the vMF concentration) denotes the scale treated as a constant during back-propagation, so the β-weight (β=0.5) re-scales gradients without itself receiving them. Standard Gaussian negative log-likelihood (NLL) permits variance collapse on easy samples. Beta-NLL [18] re-weights gradients to prevent this, where $\sigma_{\text{det}} = \text{stopgrad}(\sigma)$ denotes σ treated as a constant (no gradient propagation). An identical scalar Beta-NLL is used for the jaw angle.

$$\mathcal{L}_{\text{pos}} = \mathcal{L}_{\text{jaw}} = \tfrac{1}{2}\left[\sigma_{\text{det}}^{2\beta}\,\frac{(\mathbf{y}-\boldsymbol{\mu})^2}{\sigma^2} + \sigma_{\text{det}}^{2(1-\beta)}\log\sigma^2\right], \quad \beta{=}0.5\,, \tag{2}$$

Eq 3: Beta-vMF. Quaternion orientation on $\mathbb{S}^3$ is modelled by a von Mises–Fisher distribution with analogous beta-weighting, where $\mathbf{q}^*$ is the ground-truth quaternion and $C_4(\kappa) = \kappa^2/(4\pi^2\sinh\kappa)$ is the vMF normalising constant on $\mathbb{S}^3$.

$$\mathcal{L}_{\text{rot}} = \kappa_{\text{det}}^{\beta}\big(-|\boldsymbol{\mu}_q\cdot\mathbf{q}^*|\big) + \kappa_{\text{det}}^{1-\beta}\log C_4(\kappa)\,, \tag{3}$$

Eq 4: Differentiable ECE (Expected Calibration Error). We directly optimise uncertainty calibration, where B=9 calibration bins span target coverages $\alpha_b \in \{0.1,\ldots,0.9\}$, z_{α_b} the corresponding standard normal quantiles, σ_i the predicted standard deviation, e_i the prediction error, and τ controls temperature.

$$\mathcal{L}_{\text{ECE}} = \frac{1}{B}\sum_{b=1}^{B}(\hat{c}_b-\alpha_b)^2\,, \quad \hat{c}_b = \frac{1}{N}\sum_i \sigma\!\left(\frac{z_{\alpha_b}\sigma_i - |e_i|}{\tau\,\sigma_i}\right), \tag{4}$$

2.5 Training Strategy

The kinematic HTT is initially pre-trained for masked visual reconstruction (30% of frames masked). It will use the previously predicted kinematics (initially, the random initialisation before there are any kinematics) as priors in the HTT to produce the kinematic embeddings. The main BTPN performs visual–kinematic alignment (cosine similarity between kinematic embeddings and visual features). In fine-tuning, we first train visual projections only and then all parameters (cosine LR from 3×10^{-4} to 10^{-7}). We use AdamW (β_1=0.9, β_2=0.999, weight decay 10^{-5}), gradient clipping at 1.0, batch size 16, and dense sliding window sampling frame ($\sim$47 K training windows with stride 2). At inference, the model runs a single deterministic forward pass and reads predictive (aleatoric) uncertainty directly from the heteroscedastic output heads. All experiments ran on an NVIDIA RTX 3060 with PyTorch 2.7 and CUDA 11.8.

3 Results

3.1 Experimental Setup

We report RMSE (root mean squared error) scores between predicted and ground-truth tool-tip coordinates and overall Euclidean distance (mm), geodesic rotation [22] (shortest angular path between predicted and target quaternions; °), and percentage of the jaw opening range. Uncertainty quality is assessed by ECE (deviation of predicted confidence intervals from observed coverage rates) and empirical coverage (fraction of ground-truth poses falling within predicted 90% intervals). Visual components are scored by mAP_{50} and $\text{mAP}_{50\text{-}95}$. For cross-dataset evaluation (Table 2c), we report ΔRot, the RMSE of the per-step frame-invariant quaternion angular velocity computed between predicted and ground-truth quaternions. Due to the Bayesian nature, we also report autoregressive (AR) multi-step position measuring kinematic divergence at horizon h frames. Metrics are reported thoroughly on held-out test sets. The model is trained only on Dataset A (39 trials for training, 21 held out). No published method performs full 7-DoF monocular pose tracking, precluding direct comparison. ART-Net is the closest in setting and serves as a representative external baseline, complemented by other architectures and component ablations. In Table 2b, we compare against: **(1)** ART-Net [10]; **(2)** visual embedding with (a) regression baseline; (b) Long-Short Term Memory (LSTM) network; (c) Temporal Convolutional Network (TCN); and (d) Visual Tracking Transformer (VTT); **(3)** a kinematic regression baseline on raw EM input; **(4)** ablations removing (a) KFM priors; (b) bimanual cross-attention; (c) hierarchical multi-scale attention; and (d) uncertainty ECE loss.

3.2 Quantitative Results

BTPN substantially outperforms all other models (Table 2a). All methodological contributions provide benefits over the baseline (Table 2b). Cross-dataset evaluation (Table 2c) demonstrates generalisation.

Table 2: Quantitative results overview.

(a) Visual Pipeline Components

Component	Prec ↑	Recall ↑	mAP_{50} ↑	$mAP_{50\text{-}95}$ ↑
Detection	98.1	98.3	99.1	97.5
Segmentation	98.1	98.3	99.1	91.1
Keypoints	97.3	97.3	98.3	94.6

(b) Pose Prediction RMSE Errors (Dataset A held-out)

	Position ↓ (mm)				Rotation ↓ (°)				Jaw ↓	ECE ↓
Method	x	y	z	$\|v\|$	roll	pitch	yaw	Geo	(% open)	
ART-Net [10]	19.7	20.7	14.9	32.2	65.6	30.7	67.5	55.4	42.5	0.155
Visual regr.	18.3	17.0	13.0	28.1	52.8	27.2	55.0	44.0	45.6	0.137
Visual + LSTM	17.1	13.2	12.3	24.9	76.8	36.8	68.0	67.9	43.2	0.098
Visual + TCN	15.1	11.9	11.0	22.1	72.0	37.3	68.6	66.4	41.9	0.240
Visual + VTT	16.4	13.7	12.5	24.7	74.0	45.6	79.7	69.0	42.2	0.115
Kinematic regr.	5.2	5.8	3.9	8.7	33.1	17.2	31.2	27.6	14.4	0.098
Full BTPN	**4.2**	**4.4**	**3.4**	**7.0**	**14.4**	**7.3**	**15.6**	11.7	13.6	0.028
w/o kin. prior	17.0	13.8	12.5	25.2	77.2	33.4	75.9	55.5	39.1	0.138
w/o bimanual	5.6	5.8	4.2	9.1	20.9	8.9	22.2	13.6	14.3	0.025
w/o multiscale	4.3	4.7	3.6	7.3	18.0	7.4	19.0	**11.6**	**13.2**	**0.020**
w/o calibration	4.2	4.5	3.5	7.0	15.1	7.7	16.0	11.9	13.5	0.259

(c) Cross-Dataset Generalisation RMSE Errors (Full BTPN Model)

		Position ↓ (mm)				ΔRot ↓	Jaw ↓	AR Pos ↓	
Dataset	Role	x	y	z	$\|v\|$	(°/step)	(% open)	@5	@10
A (21 tr.)	Held-out	4.2	4.4	3.4	7.0	17.6	13.6	16.7	28.2
B (30 tr.)	In-dist.	5.6	5.0	4.2	8.6	28.9	11.6	21.9	34.9
C (24 tr.)	OOD	4.9	7.1	6.8	11.0	29.1	N/A	21.1	33.5

3.3 Qualitative Results and Uncertainty Analysis

Correlating proxy variables with position error reveals that motion speed is the strongest predictor of error ($r = 0.65$, $p < 0.0001$). Position extremity from workspace centre ($r = 0.50$) and inter-tool proximity ($r = 0.30$) are also significant ($p < 0.0001$). However, predictive uncertainty minimises these cases (Fig. 4) and we can further see representative high tracking accuracy for trajectory predictions (Fig. 3). Per-modality calibration reveals position (ECE = 0.028) and jaw angle (ECE = 0.079) are well-calibrated, while rotation is substantially over-conservative (ECE = 0.30, Fig. 4a), reflecting the inherent difficulty of recovering orientation from monocular images. Adding epistemic uncertainty from 30 Monte Carlo dropout samples [8] to the aleatoric predictions does not improve calibration: position and jaw ECE slightly worsen (to 0.13 and 0.11, from over-coverage) while rotation is unchanged (0.30), confirming that the heteroscedastic output heads already capture the predictive uncertainty. We therefore report single-pass deterministic inference.

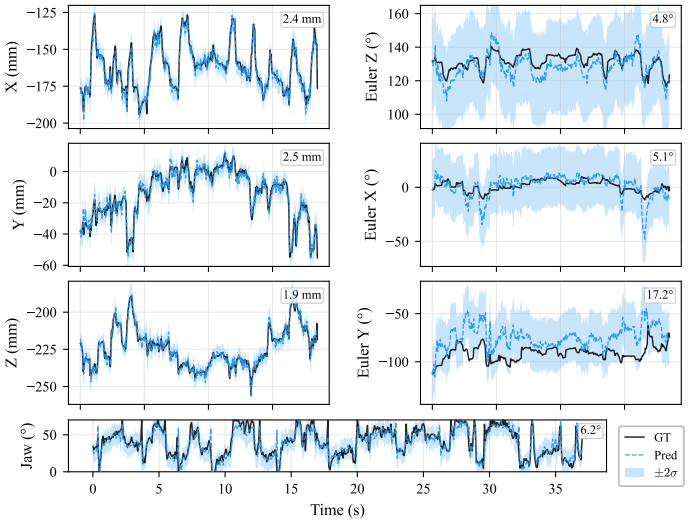

Fig. 3: 7-DoF predictions for a held-out trial sequence.

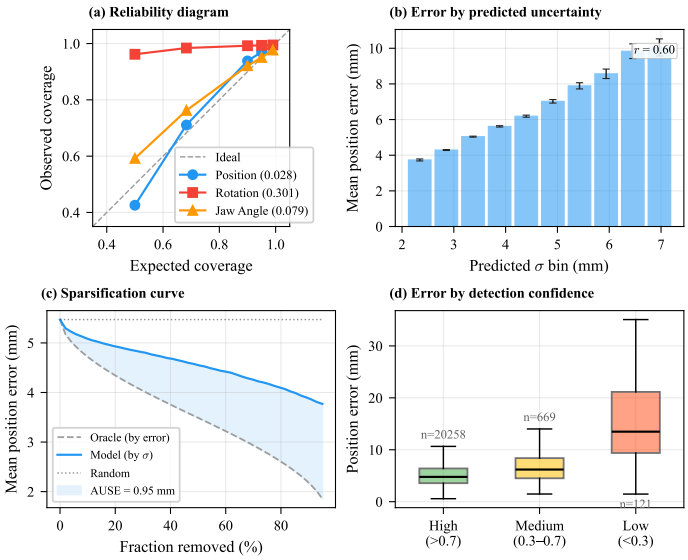

Fig. 4: Uncertainty quality assessment. (a) Reliability diagram at five confidence levels with per-DoF ECEs. (b) Mean position error confirms higher predicted aleatoric uncertainty corresponds to higher error. (c) Area Under the Sparsification curve shows the model's useful uncertainty ranking for selective prediction. (d) Position error stratified by detection confidence.

4 Discussion

Whilst this work makes several contributions, we believe the kinematic data can be incredibly valuable to researchers as it opens an entirely new, yet clinically relevant modality, especially with its potential for skill analysis. The results are clinically meaningful given the scale of laparoscopic peg transfer, sufficient to capture tool trajectories, and achieved entirely through vision. Since the in-house dataset has no comparable counterpart, it places our results as a baseline benchmark. However, the current framework has clear limitations. The models are not optimised for real-time inference on edge devices, which is required for deployment in surgical training settings. Nevertheless, these results strongly motivate future work in quantitative, objective surgical skill assessment.

Acknowledgments. We would like to thank the organisers of the Urology Simulation Boot Camps in Leeds, UK, the British Association of Paediatric Endoscopic Surgeons and all surgeons who spent time completing the data collection across all these events. The work was funded by UKRI EPSRC grants EP/S024336/1, UKRI914 and UKRI180.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ali, M., Pena, R.M.G., Ruiz, G.O., et al.: A comprehensive survey on recent deep learning-based methods applied to surgical data (2022). https://doi.org/10.48550/ARXIV.2209.01435
2. Ayobi, N., Rodríguez, S., Pérez, A., et al.: Pixel-wise recognition for holistic surgical scene understanding. Medical Image Analysis **106**, 103726 (Dec 2025). https://doi.org/10.1016/j.media.2025.103726
3. Carciumaru, T.Z., Tang, C.M., Farsi, M., et al.: Systematic review of machine learning applications using nonoptical motion tracking in surgery. npj Digital Medicine **8**(1), 1–20 (Jan 2025). https://doi.org/10.1038/s41746-024-01412-1
4. Choudhry, O., Ali, S., Biyani, C.S., Jones, D.: Real-time tool detection in laparoscopic datasets for surgical training in low-resource settings. Healthcare Technology Letters **12**(1), e70045 (2025). https://doi.org/10.1049/htl2.70045
5. Choudhry, O., Ali, S., Rajasundaram, R., Biyani, C.S., Jones, D.: 7-DoF laparoscopic peg transfer dataset for surgical skill assessment. In: Medical Image Understanding and Analysis (MIUA 2025). Frontiers Media SA (2025). https://doi.org/10.3389/978-2-8325-5137-0
6. Ding, H., Seenivasan, L., Killeen, B.D., et al.: Digital twins as a unifying framework for surgical data science: the enabling role of geometric scene understanding. Artificial Intelligence Surgery **4**(3), 109–138 (Jul 2024). https://doi.org/10.20517/ais.2024.16
7. Frisch, Y., Bornberg, C., Fuchs, M., et al.: GAUDA: Generative Adaptive Uncertainty-guided Diffusion-based Augmentation for Surgical Segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3762–3771 (Feb 2025). https://doi.org/10.1109/WACV61041.2025.00370
8. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Oct 2016). https://doi.org/10.48550/arXiv.1506.02142

9. Gautier, B., Tugal, H., Tang, B., et al.: Real-Time 3D Tracking of Laparoscopy Training Instruments for Assessment and Feedback. Frontiers in Robotics and AI **8**, 751741 (Nov 2021). https://doi.org/10.3389/frobt.2021.751741
10. Hasan, M.K., Calvet, L., Rabbani, N., et al.: Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. Medical Image Analysis **70**, 101994 (May 2021). https://doi.org/10.1016/j.media.2021.101994
11. Jocher, G., Qiu, J.: Ultralytics yolo26 (2026), https://github.com/ultralytics/ultralytics
12. Kim, T.D., Dao, N.N., Tran, D.V., et al.: Surgical Tool Detection and Pose Estimation using YOLOv8-pose Model: A Study on Clipper Tool. In: 2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV). pp. 225–229 (Jun 2024). https://doi.org/10.1109/ICDV61346.2024.10617290
13. Lambert, B., Forbes, F., Doyle, S., et al.: Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artificial Intelligence in Medicine **150**, 102830 (Apr 2024). https://doi.org/10.1016/j.artmed.2024.102830
14. Nema, S., Vachhani, L.: Surgical instrument detection and tracking technologies: Automating dataset labeling for surgical skill assessment. Frontiers in Robotics and AI **9** (Nov 2022). https://doi.org/10.3389/frobt.2022.1030846
15. Rueckert, T., Rauber, D., Maerkl, R., et al.: Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the PhaKIR 2024 challenge. Medical Image Analysis p. 103945 (2026). https://doi.org/10.1016/j.media.2026.103945
16. Rueckert, T., Rueckert, D., Palm, C.: Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art. Computers in Biology and Medicine **169**, 107929 (Feb 2024). https://doi.org/10.1016/j.compbiomed.2024.107929
17. Sangalli, S., Sarwin, G., Erdil, E., et al.: Conformal forecasting for surgical instrument trajectory (Mar 2025). https://doi.org/10.48550/arXiv.2503.04191
18. Seitzer, M., Tavakoli, A., Antic, D., et al.: On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks (Apr 2022). https://doi.org/10.48550/arXiv.2203.09168
19. Shabir, D., Padhan, J., Abdurahiman, N., et al.: A Real-Time Markerless Approach for 3D Pose Estimation of Laparoscopic Instruments in Surgical Training Simulators. IEEE Access **13**, 216646–216662 (2025). https://doi.org/10.1109/ACCESS.2025.3647882
20. Spektor, R., Friedman, T., Or, I., et al.: Monocular pose estimation of articulated open surgery tools - in the wild. Medical Image Analysis **103**, 103618 (Jul 2025). https://doi.org/10.1016/j.media.2025.103618
21. Wen, B., Yang, W., Kautz, J., et al.: FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17868–17879 (2024). https://doi.org/10.1109/CVPR52733.2024.01692
22. Wolf, R., Duchateau, J., Cinquin, P., et al.: 3D Tracking of Laparoscopic Instruments Using Statistical and Geometric Modeling. In: Fichtinger, G., Martel, A., Peters, T. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011, vol. 6891, pp. 203–210. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23623-5_26

23. Xu, H., Weld, A., Xu, C., et al.: SurgRIPE challenge: Benchmark of surgical robot instrument pose estimation. Medical Image Analysis p. 103674 (2025). https://doi.org/10.1016/j.media.2025.103674
24. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2 (Jun 2024). https://doi.org/10.48550/arXiv.2406.09414
25. Yang, S., Luo, L., Wang, Q., et al.: Surgformer: Surgical Transformer with Hierarchical Temporal Attention for Surgical Phase Recognition (Aug 2024). https://doi.org/10.48550/arXiv.2408.03867