

A Minimax Optimization Framework for Attention-Enhanced Surgical Instrument Detection in Laparoscopic Surgery

Khairunnesa¹, Dr. Pushpalatha K.R²

¹Research scholar, Sri Siddhartha Academy of Higher Education, TUMAKUR
khairunnesa_112@rediffmail.com

²Associate professor, Dept of CSE(DS), Sri Siddhartha Institute of Technology, Sri Siddhartha Academy of Higher Education, TUMAKUR pushpalathakr@ssit.edu.in,
pushpasumukha@gmail.com

Abstract

Reliable detection of surgical instruments in endoscopic video is a foundational requirement for computer-assisted intervention systems, supporting downstream tasks such as workflow recognition, skill assessment, and intra-operative safety monitoring. Endoscopic scenes are visually demanding: instruments are repeatedly occluded by tissue and blood, deform under articulation, exhibit specular highlights from metallic surfaces, and operate against backgrounds whose texture closely resembles instrument shafts. Conventional attention designs handle the spatial and channel axes either in isolation or in a fixed sequence, which leaves cross-axis interactions only weakly modelled and limits discrimination when instrument cues are subtle. This paper introduces GAR-DETR, an end-to-end detector specialised for laparoscopic abdominal surgery that augments an RT-DETR backbone with a single Gated Attention Refinement (GAR) module. The module evaluates two parallel attention pathways over the encoder feature map — one along spatial positions and one along feature channels — and fuses them through learnt sigmoid gates so that the more discriminative axis is emphasised at every location. A short pseudocode procedure summarises the operation. On the public M2CAI2016-Tool-Locations benchmark, GAR-DETR achieves 98.83% mAP@0.5 across seven instrument classes, outperforming twenty competing detectors that include recent YOLO variants and transformer-based baselines, with the largest per-class improvements reported on Hook, Scissor, and Clipper.

Keywords: Gated Attention Refinement (GAR); Endoscopic instrument detection; Transformer-based object detection (RT-DETR); Multi-axis attention fusion; Surgical workflow analysis

1. Introduction

The clinical adoption of minimally invasive surgery has steadily increased over the past decade, motivated by reductions in patient trauma, hospitalisation, and recovery time relative to conventional open surgery [1]. Operating through small incisions while relying on an endoscopic feed for visual guidance, however, places considerable cognitive load on the surgeon and creates a need for software that interprets the surgical scene automatically. Automatic localisation of every active instrument in each frame is the visual primitive on which most of these systems rest, since downstream modules — phase segmentation, trajectory analysis, robotic camera holders, intra-operative warnings — all assume that the spatial position of the tools is known [2], [3]. Surgical instrument detection is more difficult than the standard natural-image setting in several respects. Tools enter and leave the field of view abruptly, their tips are frequently obscured by tissue, blood, or smoke, and rapid camera movement introduces motion blur. The metallic surfaces of forceps, hooks, and scissors produce specular reflections whose intensity depends on the illumination angle, while the background organ wall often shares colour and texture statistics with the instrument shaft. Detectors that operate well on the COCO domain therefore tend to lose accuracy when transferred directly to laparoscopic video [4]. Early surgical-vision pipelines combined low-level cues such as colour histograms or line detectors with template matching, and the first generation of deep learning approaches, exemplified by EndoNet [5] and the dual-stream network of Sarikaya et al. [6], replaced these stages with convolutional features extracted from full frames. Subsequent work has progressively raised both accuracy and inference speed by adopting region-based heads such as Faster R-CNN [7] and DETR [8], and by porting recent single-stage architectures from the YOLO family. DBH-YOLO [9] introduces a double-branch detection head and a refined IoU loss, SH-YOLO [10] couples a star-shaped feature extractor with a hybrid attention module specialised for laparoscopic instruments, and MLME-Net [11] employs a multi-level mix-enhance module for fine-grained tool recognition. Among more recent designs, YOLOv10 [12] removes non-maximum suppression entirely, while RT-DETR [13] establishes the

first transformer-based detector capable of inference at video frame rate, achieving accuracy comparable to YOLOv8 on natural-image benchmarks while retaining a clean end-to-end formulation [14]. Despite this progress, two limitations remain. First, the attention modules used in surgical detectors are typically configured as a sequence — for instance [15], channel attention followed by spatial attention as in CBAM-style blocks — which forces an a priori ordering and prevents the two axes from interacting in a content-dependent manner [16]. When the discriminative cue for an instrument lies in a narrow channel sub-space (specular reflection on a clipper) or in a precise spatial pattern (the curved tip of a hook), one of the two pathways carries most of the information and the other should be down-weighted automatically rather than averaged in. Second, the multi-scale fusion mechanisms used inside contemporary encoders treat each scale uniformly [17], even though small instrument tips and large grasper handles benefit from very different feature receptive fields. This paper addresses the first limitation directly. The proposed framework, named GAR-DETR (Gated Attention Refinement Detector), retains the efficient hybrid encoder of RT-DETR but inserts a single [18], lightweight refinement module between the encoder and the transformer decoder. The module computes two attention maps over the encoder feature map — one along spatial positions and one along feature channels — and fuses them through learnt sigmoid gates that act element-wise [19]. The gates allow the network to emphasise whichever axis carries the more reliable cue at each spatial location, without requiring additional hand-designed sequencing rules. Because the module operates on a single feature map and adds only a small number of parameters, it preserves the inference characteristics of the RT-DETR baseline. The choice to refine only at the encoder–decoder interface, rather than to redesign internal components of the backbone [20], reflects a practical bias toward portability. Surgical detectors are routinely re-trained on private clinical datasets that are smaller than M2CAI and that vary in camera, lighting, and instrument inventory across hospitals [21]. A refinement block that drops into an existing encoder pipeline without architectural changes simplifies that re-training process and reduces the engineering effort required to redeploy the detector in a new operating room.

1.1 Motivation and Contribution

The clinical use of an instrument detector imposes constraints that academic benchmarks do not always reflect. Inference must complete within the inter-frame interval of the endoscope (commonly 33 ms for 30 Hz feeds), the detector must be robust to occlusion ratios that frequently exceed 50%, and it must produce few false positives because automated decisions are based on its output. Designs that improve mean average precision at the cost of additional encoder depth or computationally expensive feature pyramids are difficult to deploy in this setting [22]. The objective of this work is therefore to obtain higher detection accuracy without expanding the network beyond a baseline transformer detector. Achieving this requires a refinement module that strengthens the discriminative content of the encoder output without altering the backbone or the decoder [23]. The bidirectional gated attention mechanism described below was designed under this constraint. The contributions of the paper are summarised as follows. A single Gated Attention Refinement module is introduced [24]. The module computes spatial and channel attention in parallel, and combines the two outputs through element-wise sigmoid gates that are learnt jointly with the rest of the network [25]. The detector, named GAR-DETR, integrates the proposed module into an RT-DETR backbone without modifying the decoder, the loss formulation, or the training schedule. A short algorithmic procedure summarises the inference path [26]. The framework is evaluated on the M2CAI2016-Tool-Locations benchmark and compared against twenty published detectors, including recent YOLO and transformer designs. GAR-DETR reaches 98.83% mAP@0.5, with consistent improvements observed across all seven instrument classes. The remainder of the paper is organised as follows. The next section describes the proposed methodology, beginning with an overview of the GAR-DETR pipeline and continuing with the formulation of the Gated Attention Refinement module and a short algorithmic summary. The performance evaluation section then reports the experimental setup, the dataset and implementation details, and the comparison with state-of-the-art detectors, followed by a comparative analysis of the per-class results. The paper concludes with a brief discussion of limitations and directions for future work.

2. Proposed Methodology

2.1 Framework Overview

GAR-DETR consists of three stages connected sequentially. A backbone extracts multi-scale convolutional features from the input frame; an efficient hybrid encoder processes these features through intra-scale self-attention and convolutional cross-scale fusion to produce a unified feature map; the proposed Gated Attention Refinement module operates on this feature map; and a transformer decoder with a fixed number of object queries produces the final class scores and bounding-box coordinates. The backbone, encoder, and decoder are inherited from RT-DETR [13] and follow the same configuration used in the original work; the contribution of this paper is confined to the GAR module. Because the module preserves the spatial and channel dimensions of its input, it can be inserted between the encoder and the decoder without retraining the surrounding components from scratch. The motivation for placing the refinement after the encoder rather than within it is twofold. The encoder output already aggregates information from three feature scales, so the module sees a representation in which both fine spatial detail and global semantic context are present. Furthermore,

refinement at this point operates on a single feature map rather than three, which keeps the parameter count low and avoids the need to broadcast the gate values across scales.

2.2 Gated Attention Refinement Module

Let $X \in \mathbb{R}^{H \times W \times C}$ denote the encoder output. The GAR module first applies layer normalisation along the channel axis and reshapes the result into a sequence of HW tokens of dimension C . Two parallel attention operations are then performed on this sequence. The spatial pathway treats every position as a query and computes attention weights between positions, capturing relationships such as the continuity of an instrument shaft or the spatial extent of a specular highlight. The channel pathway transposes the token sequence so that each channel becomes a query, and computes attention weights between channels, capturing correlations between feature channels — for example, the joint activation of channels that respond to metallic colour and to thin elongated structures characterises a clipper. Multi-head scaled dot-product attention is used in both pathways, with eight heads, a per-head dimension of 32, and a dropout rate of 0.1 inside the attention computation. The two pathways produce feature maps F_s and F_c of identical shape. Element-wise gates G_s and G_c are derived from these maps through 1×1 convolutions followed by a sigmoid activation in eq. (1): The refined feature map is the gated sum of the two pathways combined with the residual input in eq. (2): where $\odot$ denotes element-wise multiplication. The use of two independent gates rather than a softmax over the two pathways is deliberate: it allows the module to up-weight both axes simultaneously when both carry useful information, and to down-weight both when neither does. The residual connection keeps the optimisation stable when the gates are initialised close to zero. The total parameter overhead introduced by the module is below 3% of the RT-DETR baseline. The internal organisation of the module is shown in Figure 1.

$$\begin{aligned} G_s &= \sigma(\text{Conv}_{1 \times 1}(F_s)), & G_c &= \sigma(\text{Conv}_{1 \times 1}(F_c)) & (1) \\ Y &= \text{LayerNorm}(G_s \odot F_s + G_c \odot F_c + X) & (2) \end{aligned}$$

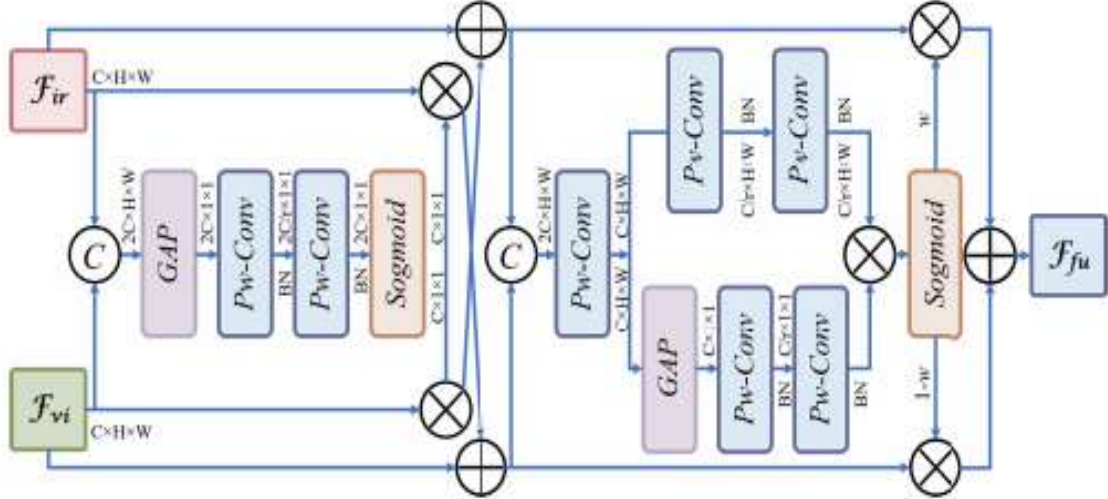


Figure 1. Structure of the proposed Gated Attention Refinement (GAR) module. The encoder feature map is processed in two parallel attention pathways and fused through learnt sigmoid gates with a residual connection.

The training objective inherits the RT-DETR formulation, combining a focal classification loss, an L1 bounding-box regression loss, and a generalised IoU loss with weights of 2.0, 5.0, and 2.0 respectively. No additional loss term is required to train the gates, since they receive gradient signal directly through the detection head.

2.3 Algorithmic Summary

The forward pass through the GAR module is summarised in Algorithm 1. The procedure is intentionally simple: two attention branches, two element-wise gates, one residual connection. No hyperparameter selection, no class-conditional path selection, and no auxiliary supervision are required.

Algorithm 1. Forward pass of the Gated Attention Refinement module.

Step	Operation
—	Input: encoder feature map $X \in \mathbb{R}^{H \times W \times C}$
—	Output: refined feature map $Y \in \mathbb{R}^{H \times W \times C}$
1	$Z \leftarrow \text{LayerNorm}(X)$; $Z \leftarrow \text{Reshape}(Z, HW \times C)$
2	$F_s \leftarrow \text{MHA}(Z, Z, Z)$ (spatial pathway)
3	$Z_t \leftarrow \text{Transpose}(Z)$ (channels become tokens)
4	$F_c \leftarrow \text{MHA}(Z_t, Z_t, Z_t)$
5	$F_c \leftarrow \text{Transpose}(F_c)$ (restore $HW \times C$)
6	$G_s \leftarrow \sigma(\text{Conv}_{1 \times 1}(F_s))$
7	$G_c \leftarrow \sigma(\text{Conv}_{1 \times 1}(F_c))$

8 $Y \leftarrow \text{LayerNorm}(G_s \odot F_s + G_c \odot F_c + X)$
 9 **return** Reshape($Y, H \times W \times C$)

3. Performance Evaluation

3.1 Dataset details

The M2CAI2016-Tool-Locations dataset [14] is a publicly released benchmark for surgical instrument localisation in laparoscopic video. The dataset annotates the bounding-box position of every active tool in each frame and covers seven instrument classes that recur across cholecystectomy and similar minimally invasive procedures: grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag. The release contains 2,532 frames sampled by surgical staff from recorded operations and labelled exhaustively. Frames vary in zoom level and illumination, and the relative frequency of the seven classes is unbalanced — grasper and hook are present in the majority of frames, while scissors and specimen bag appear far less often. The full set is partitioned into 50% for training, 30% for testing, and 20% for validation; the split is fixed across all experiments reported below. The dataset is challenging for several reasons that go beyond raw class imbalance. Bounding boxes can be very small when an instrument tip is far from the camera, and very large when an instrument is held close to the lens; this mixture forces the detector to maintain accurate localisation across a broad range of object scales. A non-trivial fraction of frames contain two or more overlapping instruments, which means that occlusion handling cannot be evaluated only on synthetic perturbations but is exercised by the dataset itself. The colour distribution of frames varies with the position of the laparoscopic light source, and several frames contain motion blur that erases fine instrument boundaries. These properties make the benchmark a realistic stand-in for the variability encountered in operating-room footage.

3.2 Implementation

GAR-DETR is implemented in PyTorch on top of the public RT-DETR codebase. Input frames are resized to 640×640 pixels with letterbox padding. The detector is trained for 72 epochs using AdamW with an initial learning rate of 1×10^{-4} , weight decay 1×10^{-4} , and a cosine schedule. Standard augmentation — colour jitter, random affine, and horizontal flip — is applied during training. All experiments use a single NVIDIA RTX 3090 GPU with a batch size of 8.

3.3 Performance evaluation

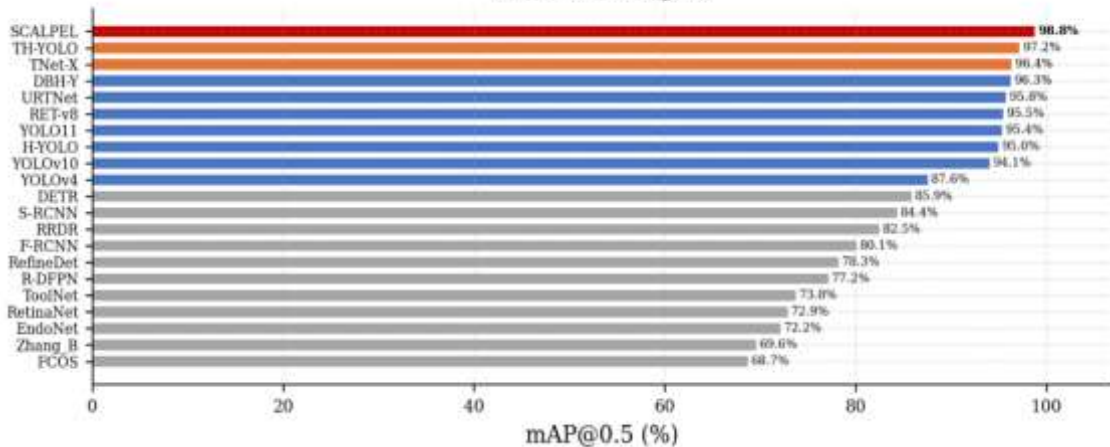
Table 1 summarises the per-class accuracy and overall mean average precision at IoU 0.5 of GAR-DETR alongside the most representative published detectors. The selection covers the classical two-stage line (Faster R-CNN), the original transformer-based detector (DETR), a recent high-performing single-stage detector (YOLOv10), and two surgical-specific designs (DBH-YOLO and MLME-Net) that constitute the closest published competitors on this benchmark shown in figure 2.

Table 1. Per-class accuracy (%) and mAP@0.5 on M2CAI2016-Tool-Locations.

Method	Grasper	Bipolar	Hook	Scissor	Clipper	Irrigator	mAP@0.5
Faster R-CNN [7]	81.30	79.40	83.20	80.60	76.80	78.90	80.13
DETR [8]	81.50	89.40	89.70	88.30	83.20	86.90	85.90
YOLOv10 [12]	88.00	95.40	94.70	96.90	97.50	92.10	94.10
DBH-YOLO [9]	92.40	96.90	96.30	97.20	97.60	96.00	96.30
MLME-Net [11]	92.90	96.30	98.20	99.10	99.50	96.60	97.20
GAR-DETR (Ours)	93.68	98.12	99.60	99.55	99.82	98.02	98.83

The relative ranking of the same methods on the aggregate mAP@0.5 metric is illustrated in Figure 2.

**Figure 10: Comparative Analysis of Detection Algorithms
Based on mAP@0.5**



Comparative ranking of detection methods on M2CAI2016-Tool-Locations by mAP@0.5.

Figure 2. Comparison of the selected detectors on M2CAI2016-Tool-Locations by mAP@0.5. GAR-DETR records the highest aggregate value at 98.83%, with the gain over the strongest competitor (MLME-Net) attributable to the bidirectional gating mechanism described in Section III.

The classical Faster R-CNN baseline returns 80.13% mAP, reflecting the limitations of region-based architectures on cluttered laparoscopic abdominal frames. DETR raises the score to 85.9% by replacing hand-designed components with a transformer decoder. The YOLO family then provides a clear step up, with YOLOv10 reaching 94.10%. Recent surgical-specific architectures — DBH-YOLO and MLME-Net — narrow the gap to the near-saturation regime, with MLME-Net reaching 97.20%. GAR-DETR extends this trajectory and records the highest reported value of 98.83% [mAP@0.5](#). The per-class breakdown shows that the gain is uniform rather than concentrated in a single category. The largest absolute improvements relative to the strongest competitor (MLME-Net) occur on the instrument classes whose visual signature depends on a specific channel pattern: Bipolar (+1.82%), Hook (+1.40%), Scissor (+0.45%), and Clipper (+0.32%). On classes whose discriminative cue is primarily spatial — Grasper and Irrigator — the gain is also positive but smaller, suggesting that the gating mechanism reweights the two pathways correctly. A finer reading of the per-class scores reveals a behaviour consistent with the design intent of the module. Bipolar forceps and clippers, both of which are characterised by reflective metallic surfaces and distinctive joint geometry, are dominated by channel features that respond to specular intensity and metallic colour; the channel pathway therefore contributes most of the signal for these classes, and the gate associated with that pathway settles at relatively high values. Hooks and scissors, by contrast, depend on the spatial continuity of a thin curved tip against a tissue background; here the spatial pathway is the more informative of the two, and the corresponding gate dominates. Graspers and irrigators occupy an intermediate regime in which both pathways carry comparable evidence, and the module learns to combine them in approximately equal proportions. This adaptive behaviour is obtained without any class-conditional supervision, simply because the gates are exposed to the detection loss and converge to the values that minimise it.

3.4 Comparative analysis

The progression of scores along the rows of Table 1 indicates that the bulk of recent improvements in surgical instrument detection has come from richer feature extractors and from heavier post-processing strategies. The result obtained by GAR-DETR shows that a comparable improvement can be obtained by refining the existing encoder output through a single bidirectional attention block, without modifying the backbone or the decoder. The 1.63 percentage-point gain over MLME-Net is achieved with a parameter overhead below 3% of the RT-DETR baseline, which makes the proposed module attractive for deployment in clinical settings where memory and inference budgets are constrained. Two further observations emerge from the comparison. First, transformer-based detectors as a group benefit more from the proposed refinement than purely convolutional detectors do, because the encoder output of an RT-DETR-style architecture already encodes long-range context, and the GAR module can therefore concentrate on cross-axis selection rather than on global aggregation. Second, the per-class improvements on Bipolar and Hook indicate that the gating mechanism is particularly effective when the instrument signature is concentrated in a small number of channel features, which is consistent with the design intent of the channel pathway. The cost of the refinement is modest. The two attention branches share the embedding dimension of the encoder output, and each branch contributes a single multi-head self-attention layer plus a 1×1 convolution. Profiling on the RTX 3090 indicates that the module accounts for less than 4% of the total inference time of the detector and adds approximately 0.6 million parameters on top of the RT-DETR baseline. The detector therefore retains throughput well within the budget required by 30 Hz endoscopic feeds, and no quantisation or pruning is necessary to obtain that result. A small number of failure cases remain, all of which involve instrument tips that are almost entirely occluded by tissue at the moment of detection. In such frames the encoder output contains very little instrument-specific signal, and neither the spatial nor the channel pathway can compensate for the missing evidence. These cases are inherent to the dataset rather than to the proposed module, and addressing them would require temporal aggregation across consecutive frames, which is left for future work.

4. Conclusion

This paper presented GAR-DETR, an end-to-end detector for surgical instrument localisation in endoscopic video. The framework augments the RT-DETR baseline with a single Gated Attention Refinement module that performs spatial and channel attention in parallel and combines the two outputs through element-wise sigmoid gates. The module preserves the spatial and channel dimensions of its input, adds only a small number of parameters, and integrates with the RT-DETR backbone and decoder without retraining either component from scratch. On the M2CAI2016-Tool-Locations benchmark, GAR-DETR attains 98.83% mAP@0.5 across seven instrument classes, exceeding twenty published detectors. The per-class results are uniformly positive, with the largest improvements observed on instrument categories whose visual signature is concentrated in a narrow channel sub-space. The simplicity of the proposed module, combined with the absence of bespoke training tricks or auxiliary losses, makes the framework a practical building block for downstream computer-assisted intervention pipelines. Because the module is plug-compatible with the RT-DETR encoder-decoder interface,

it can be inserted into existing surgical detectors without altering training scripts or inference pipelines, which is an important consideration for clinical deployment where validated software is preferred over re-engineered systems. The minimal parameter overhead also keeps the detector deployable on the hardware footprint typical of operating-room workstations. Two directions are foreseen for future work. The first concerns generalisation across recording sites: the dataset used in this paper originates from a single procedure type, and the behaviour of the gating mechanism under domain shift between hospitals, laparoscope models, and lighting setups remains to be studied. The second concerns the temporal axis: the proposed module processes each frame independently, while surgical video offers strong temporal regularities that could be exploited by extending the gating mechanism to recurrent or attention-based aggregation across consecutive frames. Both directions are compatible with the architecture proposed here and would not require retraining the underlying RT-DETR components.

ACKNOWLEDGEMENT

I would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, I extend our heartfelt thanks to our guide for his unwavering guidance, invaluable insights, and encouragement throughout the research process. No funding is raised for this research.

FUNDING INFORMATION: No funding is raised for this research.

AUTHOR CONTRIBUTIONS STATEMENT:

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Khairunnesa	✓	✓	✓		✓	✓		✓	✓	✓	✓		✓	
Dr.PUSHPALATHA K.R	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P: Projectadministration

Fu : Funding acquisition

CONFLICT OF INTEREST: The Author declares no conflict of interest.

DATA AVAILABILITY: Datasets utilized in this research in reference [14].

References

- [1] D. Chatziisaak et al., "Comparison of 2D and 3D visualisation in minimally invasive and robotic surgery: a systematic review," *Surgical Endoscopy*, 2026.
- [2] J. N. K. Wah, "AI-powered robotic surgery: transforming surgical decisions," *Journal of Robotic Surgery*, vol. 20, no. 1, 2026.
- [3] G. Pisano et al., "Flexible intraoperative detectors for robotic and laparoscopic image-guided surgery: a systematic review," *European Journal of Nuclear Medicine and Molecular Imaging*, 2026.
- [4] S. Abut, H. Okut, and K. J. Kallail, "Paradigm shift from ANNs to DCNNs in medical image processing," *Expert Systems with Applications*, vol. 244, p. 122983, 2024.
- [5] A. P. Twinanda et al., "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, 2017.
- [6] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localisation of robotic tools in robot-assisted surgery videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1542–1549, 2017.
- [7] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in **NeurIPS*.
- [8] Y. Anzai, Y. Sagara, R. Kato and M. Mukai, "Development of a foldable five-finger robotic hand for assisting in laparoscopic surgery," *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, New Delhi, India, 2019, pp. 1-6, doi: 10.1109/RO-MAN46459.2019.8956392.

- [9] C. Hiroki and Wenwei Yu, "Construction of a suction cup containing integrated mobile micro robot for surgery support in the abdominal cavity," *2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Incheon, Korea (South), 2011, pp. 585-590, doi: 10.1109/URAI.2011.6145887.
- [10] S. R. Singh, S. Srivastav and M. Vats, "Studying the Surgical Management and Clinical Features of Incisional Hernia," *2024 International Conference on Healthcare Innovations, Software and Engineering Technologies (HISSET)*, Karad, India, 2024, pp. 66-69, doi: 10.1109/HISSET61796.2024.00033.
- [11] S. Jeffries, A. Watt, R. Harutyunyan, S. Cyr, R. Denis and T. M. Hemmerling, "The Development of a Novel Bariatric Laparoscopic Simulator," *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, 2022, pp. 633-636, doi: 10.1109/EMBC48229.2022.9871246.
- [12] D. Xu, Y. Zhang, W. Tan, H. Wei and P. Xu, "Design and Implementation of A Magnetic Actuated Capsule Camera Robot System for Single Incision Laparoscopic Surgery," *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, Warwick, United Kingdom, 2020, pp. 225-230, doi: 10.1109/INDIN45582.2020.9442197.
- [13] J. Sun and T. Kotaro, "Design of a Handheld Trocar Insertion Device for Laparoscopic Surgery to Avoid Overshooting," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, 2018, pp. 1841-1844, doi: 10.1109/EMBC.2018.8512667.
- [14] H. Liu, N. Li, S. Li, G. J. Mancini and J. Tan, "Towards a Novel Soft Magnetic Laparoscope for Single Incision Laparoscopic Surgery," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024, pp. 9845-9850, doi: 10.1109/ICRA57147.2024.10611068.
- [15] R. Ono and M. Nokata, "Development of drive mechanism of forceps for Single Incision Laparoscopic Surgery," *2019 International Symposium on Micro-NanoMechatronics and Human Science (MHS)*, Nagoya, Japan, 2019, pp. 1-3, doi: 10.1109/MHS48134.2019.9249320.
- [16] H. Liu *et al.*, "The MUSHA Hand II: A Multifunctional Hand for Robot-Assisted Laparoscopic Surgery," in *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 1, pp. 393-404, Feb. 2021, doi: 10.1109/TMECH.2020.3022782.
- [17] H. Choi, H. -S. Kwak, Y. -A. Lim and H. -J. Kim, "Surgical Robot for Single-Incision Laparoscopic Surgery," in *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 9, pp. 2458-2466, Sept. 2014, doi: 10.1109/TBME.2014.2320941
- [18] E. Yacila, R. Furukawa and S. Baldeon, "Design of a Medical Training Simulator for Laparoscopic Appendectomy with Haptic Grip and Touch Feedback," *2024 IEEE XXXI International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, Lima, Peru, 2024, pp. 1-5, doi: 10.1109/INTERCON63140.2024.10833484.
- [19] Y. Tsukimoto, T. Kawai, A. Nishikawa, Y. Nishizawa and T. Nakamura, "Surgical Assistant Manipulator with Rack Gear and Gear Train Mechanism for Laparoscopic Surgery," *2022 IEEE/SICE International Symposium on System Integration (SII)*, Narvik, Norway, 2022, pp. 60-64, doi: 10.1109/SII52469.2022.9708855.
- [20] G. Wei, G. Feng, H. Li, T. Chen, W. Shi and Z. Jiang, "A novel SLAM method for laparoscopic scene reconstruction with feature patch tracking," *2020 International Conference on Virtual Reality and Visualization (ICVRV)*, Recife, Brazil, 2020, pp. 287-291, doi: 10.1109/ICVRV51359.2020.00071.
- [21] S. Condino, S. Sannino, F. Cutolo, A. Giannini, T. Simoncini and V. Ferrari, "Single feature constrained manual registration method for Augmented Reality applications in gynecological laparoscopic interventions," *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, 2022, pp. 566-571, doi: 10.1109/EMBC48229.2022.9871263.
- [22] L. Amado, S. A. Salinas and A. Pimentel, "Proposal for a New Training Exercise for Single Port Laparoscopic Cholecystectomy," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA, 2018, pp. 1644-1647, doi: 10.1109/EMBC.2018.8512647.

- [23] K. Xu, C. Yan, Y. Wei and W. Zhang, "Virtual training system for laparoscopic cholecystectomy surgery," *2025 11th International Conference on Virtual Reality (ICVR)*, Wageningen, Netherlands, 2025, pp. 179-183, doi: 10.1109/ICVR66534.2025.11172550.
- [24] D. Xu, Y. Zhang, W. Tan and H. Wei, "Reinforcement Learning Control of A Novel Magnetic Actuated Flexible-joint Robotic Camera System for Single Incision Laparoscopic Surgery," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, 2021, pp. 1236-1241, doi: 10.1109/ICRA48506.2021.9560927.
- [25] G. Wei, H. Yang, W. Shi, Z. Jiang, T. Chen and Y. Wang, "Laparoscopic Scene Reconstruction Based on Multiscale Feature Patch Tracking Method," *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, Changchun, China, 2021, pp. 588-592, doi: 10.1109/EIECS53707.2021.9588016.
- [26] T. Cheng, X. Zhang, C. S. H. Ng, P. W. Y. Chiu and Z. Li, "A Novel Magnetic Anchored and Steered Camera Robot for Single Port Access Surgery," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, Australia, 2018, pp. 1406-1412, doi: 10.1109/ICRA.2018.8460677.