MID-POSE: Multi-Instrument Detection and Pose Estimation in Endoscopic Surgery

Wenhua Wei[*1] WENHUA.WEI.17@ALUMNI.UCL.AC.UK
Laurent Mennillo[*1,2,3] L.MENNILLO@UCL.AC.UK
Zhehua Mao[1,2] Z.MAO@UCL.AC.UK
Anjana Wijekoon[1,2] A.WIJEKOON@UCL.AC.UK
Kendall Feeny[1,2] K.FEENY@UCL.AC.UK
Danyal Zaman Khan[1,2] D.KHAN@UCL.AC.UK
Evangelos B. Mazomenos[2,3] E.MAZOMENOS@UCL.AC.UK
Danail Stoyanov[1,2] DANAIL.STOYANOV@UCL.AC.UK
Hani J. Marcus[2,4] H.MARCUS@UCL.AC.UK
Sophia Bano[1,2] SOPHIA.BANO@UCL.AC.UK

[1] *Department of Computer Science, University College London, London, United Kingdom*

[2] *UCL Hawkes Institute, University College London, London, United Kingdom*

[3] *Department of Medical Physics & Biomedical Engineering, University College London, London, United Kingdom*

[4] *Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, United Kingdom*

Editors: Under Review for MIDL 2026

Abstract

Reliable perception of surgical instruments is a key prerequisite for intraoperative guidance, context-aware assistance, and workflow analysis in minimally invasive surgery (MIS). This is particularly challenging in skull base procedures, where narrow anatomical corridors, frequent occlusions, specular highlights, and visually similar instruments make multi-class detection and 2D pose estimation difficult. We address joint instrument detection and keypoint-based pose estimation from monocular endoscopic videos and introduce MID-POSE, a dual-head architecture that couples a high-resolution HRNetV2p encoder with a class-agnostic dense detection–pose head and and a Multi-level Instrument Classification (MIC) head which operates on RoI-pooled multi-level features. To support this task, we construct the PitSurg dataset from 26 clinical procedures, providing seven instrument classes with bounding boxes and detailed 2D keypoints. Using YOLOv8x-pose as our strongest baseline, which in our tasks outperforms YOLO11x-pose, MID-POSE improves Det/Pose AP$_{50–95}$ on PitSurg from 59.4/63.1 to 77.5/78.5 and on the robotic SurgPose dataset from 47.9/61.1 to 62.7/71.4. Qualitative analysis shows that high-resolution features sharpen localisation and keypoint placement, while the RoI classifier reduces misclassifications and spurious background detections, indicating that the proposed architecture and dataset provide an effective basis for robust multi-instrument perception in MIS.

Keywords: Minimally Invasive Surgery, Skull Base Surgery, Surgical Instrument Detection, Pose Estimation

[*] Contributed equally

1. Introduction

Minimally invasive surgery (MIS) has become the standard for many procedures, with surgeons operating through narrow anatomical corridors using elongated instruments visualised by an endoscope (Jeganathan et al., 2025). Joint detection and 2D pose estimation of multiple instruments offer a compact scene representation, enabling geometric reasoning and semantic understanding, which are essential for downstream applications like intraoperative guidance, context-aware assistance, workflow analysis, and skill assessment (Das et al., 2024, 2025). However, accurate instrument perception in MIS remains challenging due to issues such as specular highlights, blood, smoke, motion blur, and frequent overlap of instruments within a confined field of view. Additional complications arise from domain shifts between patients, limited dataset sizes, expensive manual annotations, and the strong visual similarity between certain tools. These factors make multi-class detection and keypoint estimation much more difficult than generic object detection or human pose estimation in natural images (Maji et al., 2022; Xu et al., 2022).

Deep learning has delivered powerful architectures for object detection and keypoint-based pose estimation (Ren et al., 2015; Newell et al., 2016; Xiao et al., 2018; Chen et al., 2018; Cheng et al., 2020; Xu et al., 2022; Maji et al., 2022). Two-stage *top-down* pipelines typically combine a generic instance detector such as Faster R-CNN (Ren et al., 2015) with a single-instance pose network (Newell et al., 2016; Xiao et al., 2018; Chen et al., 2018; Xu et al., 2022), whereas *bottom-up* methods localise all keypoints jointly and then group them into instances (Cao et al., 2017; Newell et al., 2017; Papandreou et al., 2018; Cheng et al., 2020). More recently, one-stage dense predictors such as YOLO-pose (Maji et al., 2022) jointly output boxes, classes, and keypoints in a single head, improving robustness in crowded scenes. Most of these approaches rely on backbones that repeatedly downsample the input and then attempt to recover spatial detail, whereas high-resolution representations are crucial for localisation-sensitive tasks (Sun et al., 2019).

These generic architectures have increasingly been adapted to surgical instruments. On robotic MIS, Wu et al. (Wu et al., 2025) benchmark YOLOv8x-pose, ViTPose, and DeepLabCut on the SurgPose dataset, showing that human-pose architectures can be transferred to articulated tools. For manual laparoscopy, teams in the PhaKIR challenge (Rueckert et al., 2025) extend YOLOv8-based detectors to predict per-instrument keypoints, including strategies for uncertainty estimation and handling a variable number of keypoints per class. Other works target 6D instrument pose from monocular images using one-stage regression (Yoshimura et al., 2020) or two-stage pipelines that combine YOLO-based detection with crop-based pose networks (Spektor et al., 2024). These studies demonstrate that YOLO-style one-stage detector–pose architectures are strong backbones for surgical instruments and that high-resolution encoders further benefit keypoint localisation. However, support for *truly multi-class, multi-instrument 2D pose estimation from monocular endoscopic views* remains limited, particularly in complex skull base surgery.

Dataset availability is a further bottleneck. SurgPose (Wu et al., 2025) provides articulated 2D keypoints for six robotic instrument types in stereo MIS, and PhaKIR (Rueckert et al., 2025) offers 2D keypoints for 19 laparoscopic instruments, while ROBUST-MIPS (Han et al., 2025) and ART-Net (Hasan et al., 2021) focus on tip and shaft representations for abdominal or robotic procedures. These resources provide valuable benchmarks for robotic and abdominal laparoscopy, but they do not cover monocular endoscopic pituitary surgery, where narrow corridors, strong specularities, and frequent occlusions are common, and where instruments such as pituitary rongeurs and cup forceps are visually similar and often truncated. To the best of our knowledge, there is no public dataset for endoscopic pituitary surgery that combines multi-class instrument labels with detailed, instrument-specific 2D keypoint annotations under the severe occlusion patterns encountered in the sellar phase (Marcus et al., 2021).

This work addresses both methodological and dataset gaps by proposing MID-POSE (Multi-Instrument Detection and Pose Estimation in Endoscopic surgery), a dual-head architecture for multi-class surgical instrument detection and 2D keypoint pose estimation in MIS, and by constructing a new dataset for endoscopic pituitary surgery. MID-POSE combines a high-resolution HRNet (Sun et al., 2019) encoder with a class-agnostic dense detection–pose head in the style of YOLOv8-pose (Maji et al., 2022) and a proposed Multi-level Instrument Classification (MIC) head which operates on RoI-pooled multi-level features. We evaluate the approach on both the new PitSurg dataset for manual endoscopic pituitary surgery and on the robotic SurgPose benchmark (Wu et al., 2025), allowing us to assess performance across manual and robotic MIS scenarios. The main contributions of this work are:

- **PitSurg: a dataset for endoscopic pituitary surgery**, comprising monocular intraoperative images from 26 procedures with seven instrument types annotated by bounding boxes and detailed, class-specific 2D keypoints under frequent occlusion, truncation, and class imbalance.
- **MID-POSE: a dual-head architecture for multi-class instrument detection and 2D keypoint pose estimation**, which builds on HRNetV2p features with a class-agnostic dense detection–pose head and a MIC head, and incorporates a quality-aware instrumentness objective together with an extended keypoint visibility scheme.
- **A comprehensive evaluation on PitSurg and SurgPose** that compares single-head and dual-head designs with YOLOv8/YOLO11 and HRNetV2p encoders, and demonstrates that MID-POSE consistently improves Det/Pose AP$_{50–95}$ over strong YOLOv8x-pose baselines, particularly for visually similar instruments and under occlusion in both manual and robotic MIS.

2. Method: MID-POSE for Instrument Detection and Pose

We propose MID-POSE, a dual-head architecture for multi-class surgical instrument detection and 2D keypoint pose estimation in minimally invasive surgery. The model combines a high-resolution encoder, a class-agnostic dense detection–pose head, and a MIC head operating on RoI-pooled multi-scale features (Fig. 1).

2.1. Architecture Overview

Encoder and feature pyramid: We adopt HRNetV2p-W32 as the encoder (Sun et al., 2019) as it maintains a high-resolution stream fused with lower-resolution branches, yielding semantically rich, spatially precise features that have proven effective for localization-sensitive dense prediction tasks. The encoder outputs a three-level feature pyramid (P_3, P_4,

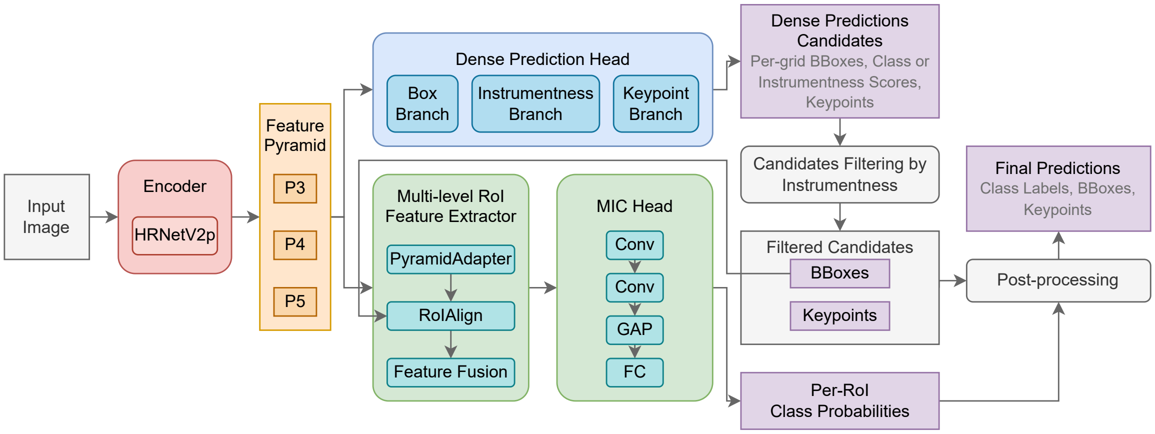

Figure 1: Overview of the proposed MID-POSE architecture for instrument detection and 2D keypoint estimation. Input images are encoded using an HRNetV2p encoder to produce a three-level feature pyramid (P3–P5). A class-agnostic dense prediction head attached to the pyramid outputs per-grid bounding boxes, instrumentness scores, and keypoints candidates that are then filtered by instrumentness. Filtered candidates are then used to pool multi-level RoI features from P3–P5 via a PyramidAdapter and RoIAlign, followed by a MIC head to predict per-RoI instrument class probabilities. These are finally combined with the dense predictions during post-processing to obtain the instrument detections and 2D keypoints.

and P_5) with strides 8, 16, and 32. These feature maps preserve fine spatial detail while progressively enriching semantics, and are shared between the dense and the MIC heads.

Class-Agnostic Dense Head: On top of P_3–P_5, we adopt a YOLOv8-Pose style head (Maji et al., 2022) for joint binary detection and 2D pose estimation in a class-agnostic manner. At every spatial location a on each pyramid level, the head predicts: (i) an instrumentness score $p_a = \sigma(z_a)$, where $\sigma(\cdot)$ denotes the sigmoid function and z_a is the corresponding raw logit, (ii) a bounding box $b_a = (x_{a,1}, y_{a,1}, x_{a,2}, y_{a,2})$, and (iii) K keypoints $k_{a,i} = (x_{a,i}, y_{a,i}, p_{v,a,i})$, $i = 1, \ldots, K$, with 2D coordinates $(x_{a,i}, y_{a,i})$ and visibility probabilities $p_{v,a,i} \in [0, 1]$. Predictions with instrumentness below a fixed threshold τ are discarded. The remaining candidates are used both as final detection/pose outputs and as proposals for the multi-level RoI classification head.

Multi-scale RoI Feature Extractor: To assign instrument categories, we introduce a multi-level RoI feature extractor operating on P_3–P_5. The bounding boxes of the filtered candidates are treated as regions of interest (RoIs). A PyramidAdapter first maps each feature map P_ℓ to a common channel dimension C_r via a 1×1 convolution,

$$P_3, P_4, P_5 \to \tilde{P}_3, \tilde{P}_4, \tilde{P}_5 \in \mathbb{R}^{C_r \times H_\ell \times W_\ell},$$

while preserving their original spatial resolutions. For every RoI, RoIAlign is applied independently to $\tilde{P}_3, \tilde{P}_4, \tilde{P}_5$, yielding tensors of size $C_r \times H_{\text{roi}} \times W_{\text{roi}}$ at each level. These three tensors are concatenated along the channel dimension to form a multi-scale descriptor of shape $3C_r \times H_{\text{roi}} \times W_{\text{roi}}$ that combines detailed local information from P_3 with increasingly contextual features from P_4 and P_5.

The concatenated tensor is passed through a feature fusion module consisting of a 1×1 convolution, group normalization, and a SiLU activation, which reduces the channels back to C_r and learns to mix information across levels. The output is a fused RoI feature map for each candidate bounding box.

MIC head: The MIC head operates on the fused RoI features and predicts a discrete probability distribution over the instrument categories. It consists of two convolutional refinement blocks (Conv), each composed of a 3×3 convolution with stride 1 and padding 1, followed by group normalization and a SiLU activation; these blocks refine the local RoI features while preserving the $H \times W$ spatial resolution of the feature map. Global average pooling (GAP) is then applied over the spatial dimensions to obtain a C_r-dimensional descriptor for each RoI. This descriptor is passed through a fully connected block (FC) with two layers, where the first is a hidden linear layer with ReLU activation and dropout and the second is an output layer that produces eight logits corresponding to the seven pituitary instrument types and a background class.

2.2. Loss Functions

Instrumentness loss: Let $y_a \in [0, 1]$ be a soft target that reflects how well the predicted instrument at location a matches its ground truth, increasing with both the instrumentness confidence and the IoU between the predicted and ground-truth bounding boxes. A quality-aware focal loss inspired by VarifocalNet (Zhang et al., 2021) is used to train this branch. The per-location instrumentness loss is defined as

$$\mathcal{L}_{\text{inst}}(p_a, y_a) = w(p_a, y_a) \left[-y_a \log p_a - (1 - y_a) \log(1 - p_a) \right], \tag{1}$$

with

$$w(p_a, y_a) = \mathbb{1}[y_a > 0]\, y_a + \mathbb{1}[y_a = 0]\, \alpha\, p_a^{\gamma}, \tag{2}$$

where α and γ control the balance between positive and negative locations and the degree of focusing on hard negatives. Because the dense head produces predictions at every spatial location while only a small top-k subset is assigned as positives, this quality-aware focal reweighting prevents the loss from being dominated by easy background locations and encourages the model to focus on well-localized positives and hard negatives.

Bounding box loss: The bounding box loss $\mathcal{L}_{\text{box}}$ follows the default YOLOv8-Pose formulation (Maji et al., 2022) and is computed only for positive grid locations ($y_a > 0$).

Keypoint loss: The keypoint loss $\mathcal{L}_{\text{kpt}}$ also follows the YOLOv8-Pose formulation (Maji et al., 2022) and is computed only for positive locations ($y_a > 0$), with separate coordinate and visibility terms $\mathcal{L}_{\text{OKS}}$ and $\mathcal{L}_{\text{vis}}$. Originally, each keypoint j has a visibility label $v_j \in \{0, 1, 2\}$ (not visible, occluded, visible). We extend this to $v_j \in \{-1, 0, 1, 2\}$ by introducing $v_j = -1$ for unannotated keypoints. For keypoints with $v_j = -1$ we set both $\mathcal{L}_{\text{OKS}}$ and $\mathcal{L}_{\text{vis}}$ to 0. This allows instances without keypoint annotations to still supervise detection, without their missing keypoints corrupting the pose supervision.

MIC Classification loss: For each RoI r, the MIC head outputs logits $s_r \in \mathbb{R}^8$ over seven instrument classes plus background. Let $y_r \in \{0, \ldots, 7\}$ denote the ground-truth label for RoI r and $p_{r,c}$ the softmax probability for class c. The MIC loss is the standard cross-entropy

$$\mathcal{L}_{\text{mic}} = \frac{1}{N_{\text{mic}}} \sum_{r=1}^{N_{\text{mic}}} \left[- \log p_{r,y_r} \right], \tag{3}$$

where N_{mic} is the number of RoIs in the batch. This penalizes low predicted probability for the ground-truth class and drives the RoI head to discriminate between the seven instrument categories and background.

Total loss. The network is trained end-to-end by combining the dense detection–pose losses from the class-agnostic head with the categorical loss from the MIC head. For each batch, the total loss is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{box}} \mathcal{L}_{\text{box}} + \lambda_{\text{inst}} \mathcal{L}_{\text{inst}} + \lambda_{\text{kpt}} \mathcal{L}_{\text{kpt}} + \lambda_{\text{mic}} \mathcal{L}_{\text{mic}} \, , \tag{4}$$

where $\lambda_{\text{box}}, \lambda_{\text{inst}}, \lambda_{\text{kpt}}$, and λ_{mic} control the relative contributions of localization, instrumentness, pose, and MIC.

3. Dataset and Experimental Setup

We evaluate MID-POSE on two complementary datasets: *PitSurg*, a new dataset of manual endoscopic pituitary surgery, and *SurgPose*, a public benchmark of robotic MIS.

PitSurg dataset is derived from 26 videos of monocular endoscopic pituitary surgery performed at the National Hospital for Neurology and Neurosurgery. We use only sellar-phase frames (Marcus et al., 2021) that contain one or two visible instruments, and split the data at the procedure level so that all frames from a given surgery appear in either the training or validation set, avoiding patient-level leakage. Seven instrument types are annotated, as illustrated in Fig. 2, namely Suction, Dural Scissors, Kerrison Rongeurs, Retractable Knife, Ring Curette, Pituitary Rongeurs, and Cup Forceps. Each instance has a bounding box, class label, and 2D keypoints with class-specific layouts, with suction having 2 keypoints, ring curette 3, retractable knife, dural scissors, pituitary rongeurs, and cup forceps 4, and Kerrison rongeurs 5. The dataset reflects real intraoperative conditions with frequent overlap, motion blur, and partial occlusion, so not all keypoints are visible in every frame. Only suction can appear together with any other instrument. Because clinically verified cup forceps annotations are scarce in the sellar phase, we augment the training set with 376 additional cup forceps instances annotated only with bounding boxes and class labels, extracted from non-sellar segments of the same procedures, while keeping the validation set restricted to sellar-phase frames. The training split contains 1,042 suction, 351 dural scissors, 495 Kerrison rongeur, 331 retractable knife, 312 ring curette, 301 pituitary rongeur, and 487 cup forceps instances, and the validation split contains 201, 80, 114, 95, 91, 51, and 83 instances, respectively.

SurgPose dataset (Wu et al., 2025) contains stereo endoscopic videos acquired with a da Vinci surgical system. Following the official protocol, we use only the left-view images and

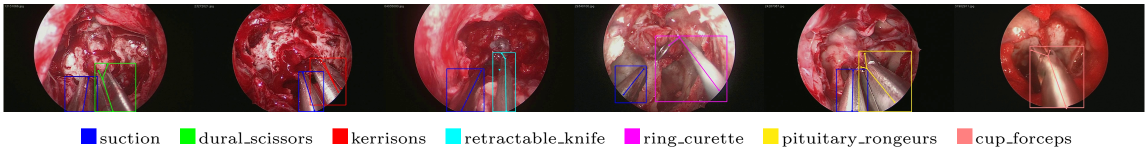

Figure 2: Examples of PitSurg instrument classes with bounding boxes and 2D keypoints.

adopt the provided split, using trajectories 0–19 for training and 20–33 for validation. Each frame contains two articulated robotic instruments annotated with bounding boxes and five 2D keypoints per tool across six instrument types: Large Needle Driver (LND), Mega Needle Driver (MND), MicroForceps, Curved Scissor, DeBakey Forceps, and Prograsp Forceps.

Five model variants are considered in our experiments, namely YOLOv8x-pose (Maji et al., 2022) and YOLO11x-pose (Ultralytics, 2024); a YOLOv8x-pose+MIC, where the YOLOv8x-pose is augmented with the proposed MIC head to form a dual-head architecture; an HRNetV2p-pose model, which uses an HRNetV2p encoder feeding a YOLO-style dense detection–pose head; and the proposed MID-POSE architecture.

Training protocol: All variants are implemented in PyTorch using Ultralytics YOLO, with custom extensions for MID-POSE, initialised from COCO-pretrained checkpoints and trained on a single NVIDIA GeForce RTX 4090 GPU. For both datasets, all models use the same augmentations. Images are resized to 640 × 640 and augmented with random rotations (up to ±20°), translations (up to 10%), isotropic scaling in $[0.5, 1.5]$, horizontal flips (probability 0.5), and mild photometric jitter in hue, saturation, and brightness.
Optimisation uses stochastic gradient descent (SGD) with a learning rate of 0.01, momentum 0.9, and weight decay 5×10^{-4}. Models are trained for 80 epochs on PitSurg and 50 epochs on SurgPose, with a batch size of 16 for YOLOv8x-pose, YOLO11x-pose, and YOLOv8x-pos+MIC, and 8 for HRNetV2p-Pose and MID-POSE. YOLOv8x-pose+MIC and MID-POSE use two-stage training: first the encoder and class-agnostic dense head are trained as a binary detector–pose model, then the full dual-head architecture is fine-tuned while RoIs are constructed online from dense predictions. For each ground-truth instrument we keep the three predictions with the highest p_a as positive RoIs and three hard negatives as the highest-p_a predictions with zero IoU to all ground-truth boxes, labelling them with the corresponding instrument class or background. At inference, dual-head models filter dense predictions with an instrumentness threshold $\tau = 0.3$, keeping only predictions with $p_a \geq \tau$ as candidates for MIC.
Following Eq. (4), all λ parameters were chosen empirically. We set $\lambda_{\text{box}} = 7.5$ and $\lambda_{\text{kpt}} = 12.0$ for all variants on both datasets. On PitSurg we use $\lambda_{\text{cls}} = 1.0$ for the single-head variants (YOLOv8x-Pose, YOLO11x-Pose, HRNetV2p-Pose), where $\mathcal{L}_{\text{cls}}$ is the multi-class counterpart of $\mathcal{L}_{\text{inst}}$. For the dual-head variants (YOLOv8x-pose+MIC and MID-POSE) on PitSurg we use $\lambda_{\text{inst}} = 1.0$ and $\lambda_{\text{mic}} = 10.0$. On SurgPose, single-head models use $\lambda_{\text{cls}} = 40.0$, and dual-head models use $\lambda_{\text{inst}} = 1.0$ and $\lambda_{\text{mic}} = 17.0$. In all cases, the λ values are set according to the relative scales and difficulty of the underlying tasks so that each loss contributes in a balanced way.

Table 1: Detection and pose $\text{AP}_{50\text{–}95}$ (in %) on the (a) PitSurg and (b) SurgPose validation sets for different encoder–head combinations. Results are reported per instrument class and as an overall class-averaged AP for detection (Det) and pose (Pose).

Architecture	(a) PitSurg Dataset - $\text{AP}_{50\text{–}95}$ (Det/Pose)							
	Overall	suction	dural_scissors	kerrisons	retractable_knife	ring_curette	pituitary_rongeurs	cup_forceps
YOLOv8x-pose (Maji et al., 2022)	59.4 / 63.1	70.6 / 84.5	65.1 / 79.5	81.6 / 57.5	53.5 / 49.6	37.7 / 66.4	56.3 / 57.6	50.8 / 46.3
YOLO11-pose (Ultralytics, 2024)	54.8 / 62.0	66.5 / 85.2	56.7 / 74.2	75.9 / 59.1	42.2 / 36.7	45.1 / 72.9	51.2 / 57.1	45.7 / 49.1
YOLOv8x-pose+MIC	59.6 / 64.9	74.7 / 87.7	57.5 / 73.4	76.9 / 67.6	57.8 / 54.1	48.9 / 76.6	50.1 / 44.3	51.6 / 51.0
HRNetV2p-pose	73.3 / 76.5	80.0 / **92.5**	72.1 / **82.3**	88.2 / 77.4	70.6 / 64.2	63.2 / 82.4	63.1 / 66.1	**75.6** / 70.4
MID-POSE	**77.5** / **78.5**	**82.9** / 92.3	**76.1** / **82.3**	**89.6** / **81.2**	**78.7** / **65.6**	**73.6** / **86.8**	**68.3** / **68.2**	73.3 / **73.2**

Architecture	(b) SurgPose Dataset - $\text{AP}_{50\text{–}95}$ (Det/Pose)						
	Overall	LND	MND	MicroForceps	Scissor	Forceps	Prograsp
YOLOv8x-pose (Maji et al., 2022)	47.9 / 61.1	74.2 / 82.0	31.4 / 30.6	26.4 / 36.2	42.4 / 83.9	32.7 / 41.4	**80.2** / **92.6**
YOLO11x-pose (Ultralytics, 2024)	47.2 / 60.9	72.3 / 80.8	31.3 / 30.4	26.0 / 36.2	41.8 / 83.8	32.2 / 41.5	79.8 / 92.6
YOLOv8x-pose+MIC	59.7 / 68.2	77.6 / 79.3	61.9 / 54.5	57.3 / 66.0	54.0 / **95.0**	39.1 / 43.7	68.3 / 70.9
HRNetV2p-pose	50.3 / 64.9	76.0 / **84.7**	41.9 / 45.9	47.6 / 55.6	26.9 / 66.0	36.5 / 48.8	72.9 / 88.3
MID-POSE	**62.7** / **71.4**	**79.5** / 81.2	**63.4** / **55.9**	**59.5** / **68.5**	**55.2** / **95.0**	**48.8** / **56.0**	69.6 / 72.1

Evaluation metrics: We report detection and pose performance using average precision over thresholds from 0.5 to 0.95 for IoU (Det $\text{mAP}_{50\text{–}95}$) and OKS (Pose $\text{mAP}_{50\text{–}95}$), given per class and as a class-averaged overall score in percentage. For qualitative examples, we show the mean IoU and OKS per image, assigning IoU $= 0$ and OKS $= 0$ to false negatives and computing these scores only for true positives.

4. Results and Discussion

We compare dense single-head architectures (YOLOv8x-pose, YOLO11-pose, HRNetV2p-pose) and dual-head variants (YOLOv8x-pose+MIC and MID-POSE) on the PitSurg and SurgPose datasets. Performance is measured using Det/Pose $\text{AP}_{50\text{–}95}$ per class and overall (Table 1). As YOLO11 does not yield consistent gains over YOLOv8x, we use YOLOv8x-pose as the primary baseline and include YOLO11-pose for completeness.

Quantitative Table 1 reports detection and pose $\text{AP}_{50\text{–}95}$ on the PitSurg and SurgPose datasets, respectively. On PitSurg, the primary performance driver is the encoder: replacing YOLOv8x with HRNetV2p improves baseline $\text{AP}_{50\text{–}95}$ from 59.4% / 63.1% (Det/Pose) to 73.3% / 76.5%, with particular gains on challenging classes like *retractable_knife*, *ring_curette* and *cup_forceps*. The MIC head adds a further boost, resulting in the best performance of 77.5% / 78.5%. Conversely, on SurgPose, the dual-head design acts as the dominant factor. While the HRNetV2p encoder offers modest gains over the YOLOv8x baseline (47.9% / 61.1%), adding the MIC head to YOLOv8x jumps performance to 59.7% / 68.2%. The proposed MID-POSE architecture achieves the best overall SurgPose results (62.7% / 71.4%). This indicates a complementary relationship: high-resolution features (HRNetV2p) drive spatial precision, critical for PitSurg, while the MIC head resolves class confusion among similar instrument tips (e.g., MND vs. MicroForceps), which is the primary bottleneck in SurgPose.

Qualitative Qualitative examples (Fig. 3 and Fig. 4) illustrate the distinct roles of the encoder and the MIC head. Across both datasets, the HRNetV2p encoder enhances lo-

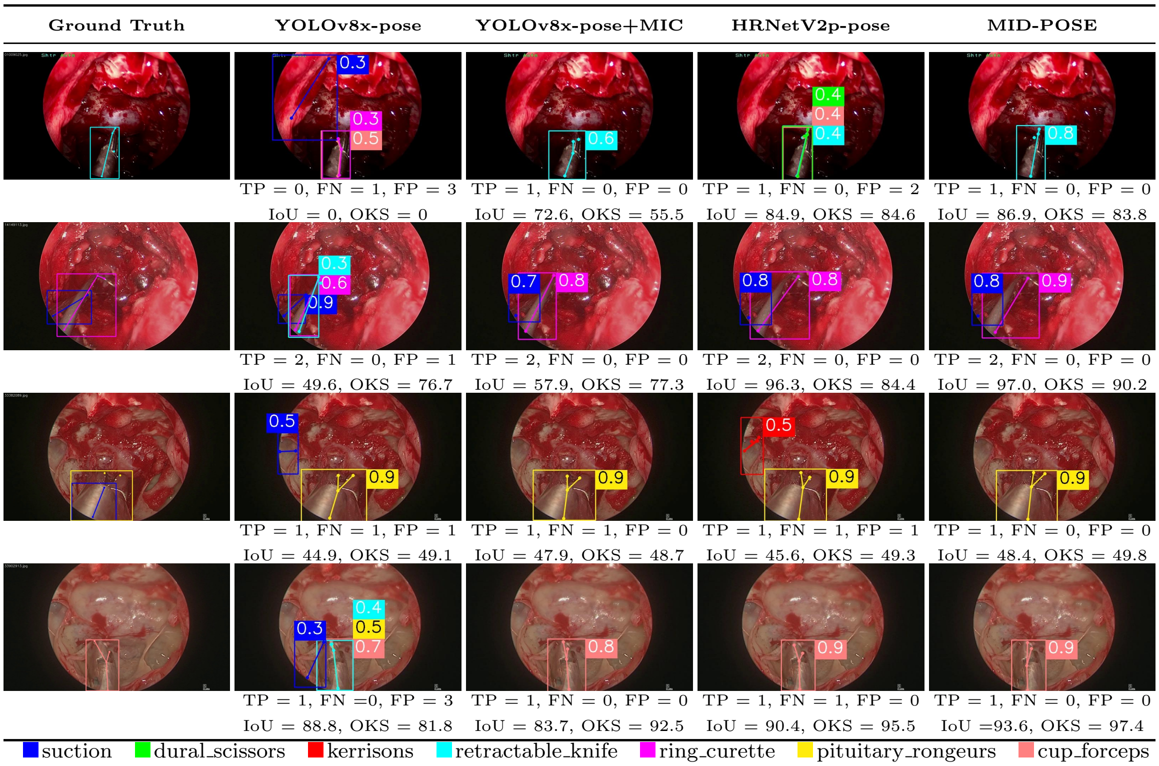

Figure 3: Qualitative PitSurg examples with ground truth and predictions. TP/FN/FP counts and mean IoU/OKS are shown below each prediction.

calization; it produces bounding boxes that tightly follow instrument shafts and captures occluded parts (e.g., *ring_curette*) where the baseline often over-extends into background tissue. In contrast, the MIC head primarily improves semantic consistency. It suppresses background false positives and corrects mislabeled bounding boxes, specifically in SurgPose, where single-head baselines struggle to distinguish instruments with similar end-effectors. While the encoder improves per-example IoU and OKS through richer spatial features, the MIC head ensures that geometrically similar bounding boxes receive coherent class labels and higher confidence scores, driving the AP gains through true positive recovery rather than further spatial refinement.

Discussion Overall, PitSurg and SurgPose highlight complementary strengths of the proposed architecture: HRNetV2p mainly improves spatial precision, whereas the MIC head addresses fine-grained semantic ambiguities between visually similar instruments. Nevertheless, several limitations remain. First, the models still exhibit persistent class confusion under extreme perspectives and strong foreshortening, as reflected by the relatively low AP for *cup_forceps* and *pituitary_rongeurs* on PitSurg and the lowest detection AP for *Forceps* on SurgPose, even with MID-POSE. Second, instruments can be missed when they are completely overlapped, or when they are very small and low contrast near the edge of the endoscopic view. For example, in Fig. 3 (row 3), all models miss the suction passing in front of the scissors. Third, keypoint localisation degrades under rare viewpoints that are

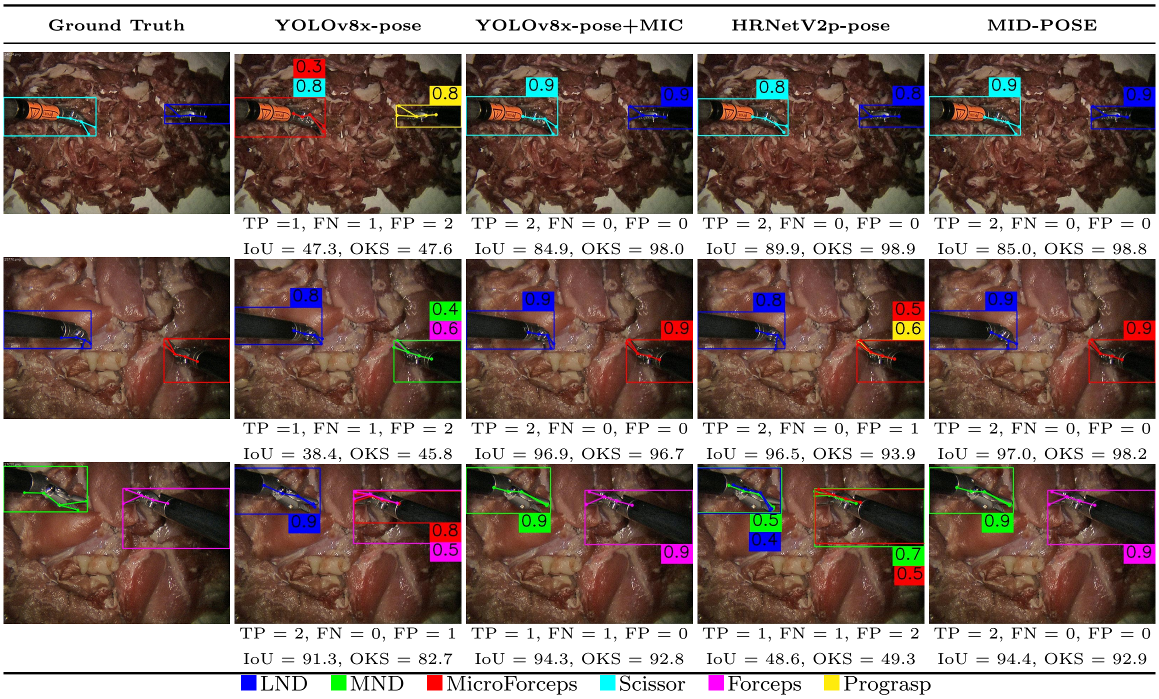

Figure 4: Qualitative SurgPose examples with ground truth and predictions. TP/FN/FP counts and mean IoU/OKS per image are reported below each prediction.

under-represented in the training set. For instance, in Fig. 4 (row 3), the tips of the MND are not correctly localised. These failure modes indicate that the proposed method is still sensitive to extreme viewpoints, heavy overlap, low contrast, and rare poses.

5. Conclusion

We presented MID-POSE, a dual-head architecture for multi-class surgical instrument detection and 2D keypoint pose estimation in minimally invasive surgery, together with the PitSurg dataset of endoscopic pituitary procedures with class-specific 2D keypoint annotations. By combining a high-resolution HRNetV2p encoder, a class-agnostic dense detection–pose head, and a MIC head operating on RoI-pooled features, MID-POSE consistently improves Det/Pose $\mathrm{AP}_{50\text{–}95}$ over strong YOLOv8x-pose baselines on both PitSurg and the robotic SurgPose benchmark, with particularly large gains for visually similar instruments and under occlusion. Qualitative results confirm that high-resolution features mainly enhance localisation and pose accuracy, whereas the MIC head resolves fine-grained class ambiguities and suppresses background false positives. Future work will explore spatio-temporal modelling of instrument appearance and motion across video frames to improve robustness, weaker forms of supervision to reduce reliance on dense annotations, and integration into real-time surgical assistance systems.

Acknowledgments

This work was supported in whole, or part, by the UCL Hawkes Institute - formerly WEISS (203145/Z/16/Z); the EPSRC (EP/Y01958X/1, EP/W00805X/1); UKRI [UKRI145] grants; DZK is supported by the NIHR Academic Clinical Fellowship; HJM is supported by the WEISS [NS/A000050/1] and by the NIHR Biomedical Research Centre at UCL, has shares in and is employed by Panda Surgical Ltd.; DS is supported by the Department of Science, Innovation and Technology (DSIT), and the Royal Academy of Engineering under the Chair in Emerging Technologies programme. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Adrito Das, Bilal Sidiqi, Laura Mennillo, Zheng Mao, Mikael Brudfors, Mihai Xochicale, Danyal Z. Khan, Nicola Newall, John G. Hanrahan, Matthew J. Clarkson, and Danail Stoyanov. Automated surgical skill assessment in endoscopic pituitary surgery using real-time instrument tracking on a high-fidelity bench-top phantom. *Healthcare Technology Letters*, 11(6):336–344, 2024.

Adrito Das, Danyal Z. Khan, Dimitrios Psychogyios, Yitong Zhang, John G. Hanrahan, Francisco Vasconcelos, You Pang, Zongyuan Chen, Jia Wu, Xiaoyun Zou, Guoyan Zheng, and Sophia Bano. Pitvis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery. *Medical Image Analysis*, page 103716, 2025. doi: 10.1016/j.media.2025.103716.

Zhe Han, Charlie Budd, Gongyu Zhang, Huanyu Tian, Christos Bergeles, and Tom Vercauteren. Robust-mips: A combined skeletal pose and instance segmentation dataset for laparoscopic surgical instruments. *arXiv preprint arXiv:2508.21096*, 2025. doi: 10.48550/arXiv.2508.21096.

Md Kamrul Hasan, Lilian Calvet, Navid Rabbani, and Adrien Bartoli. Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Medical Image Analysis*, 70:101994, 2021. doi: 10.1016/j.media.2021.101994.

J. Ravichandran Jeganathan, Ravindran Jegasothy, and Woon Teen Sia. Minimally invasive surgery: A historical and legal perspective on technological transformation. *Journal of Robotic Surgery*, 19(1):408, 2025. doi: 10.1007/s11701-025-02589-7.

Debapriya Maji, Leonardo Kispe, Gaurav Pandey, Vishnu Gandhi, Ibrahim Papandreou, and Can Wang. YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Hani J. Marcus, Danyal Z. Khan, Anouk Borg, Michael Buchfelder, Justin S. Cetas, Justin W. Collins, Neil L. Dorward, et al. Pituitary society expert delphi consensus: Operative workflow in endoscopic transsphenoidal pituitary adenoma resection. *Pituitary*, 24(6):839–853, 2021.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2016.

Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *European Conference on Computer Vision (ECCV)*, pages 282–299, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Tobias Rueckert, David Rauber, et al. Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the phakir 2024 challenge. *arXiv preprint*, arXiv:2507.16559, 2025. URL https://arxiv.org/abs/2507.16559.

Robert Spektor, Tom Friedman, Itay Or, Gil Bolotin, and Shlomi Laufer. Monocular pose estimation of articulated surgical instruments in open surgery. *arXiv preprint*, arXiv:2407.12138, 2024. URL https://arxiv.org/abs/2407.12138.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

Ultralytics. Yolo11: Next-generation real-time object detection. https://docs.ultralytics.com/models/yolo11/, 2024.

Zijian Wu, Adam Schmidt, Randy Moore, Haoying Zhou, Alexandre Banks, Peter Kazanzides, and Septimiu E. Salcudean. Surgpose: a dataset for articulated robotic surgical tool pose estimation and tracking. In *Proceedings of the IEEE International*

Conference on Robotics and Automation (ICRA), 2025. doi: 10.1109/ICRA55743.2025.11127958.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

Yan Xu, Jingbo Zhang, Qiang Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, volume 35, pages 38571–38584, 2022.

Masakazu Yoshimura, Murilo M. Marinho, Kanako Harada, and Mamoru Mitsuishi. Single-shot pose estimation of surgical robot instruments' shafts from monocular endoscopic images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9197–9203, 2020. doi: 10.1109/ICRA40945.2020.9196779.

Hongyi Zhang, Ying Wang, Feras Dayoub, and Niko Sünderhauf. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8514–8523, 2021.