Scientific Data

https://doi.org/10.1038/s41597-026-06938-5

Article in Press

ROBUST-MIPS: A Combined Skeletal Pose and Instance Segmentation Dataset for Laparoscopic Surgical Instruments

Received: 27 August 2025

Accepted: 19 February 2026

Cite this article as: Han, Z., Budd, C., Zhang, G. *et al.* ROBUST-MIPS: A Combined Skeletal Pose and Instance Segmentation Dataset for Laparoscopic Surgical Instruments. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06938-5

Zhe Han, Charlie Budd, Gongyu Zhang, Huanyu Tian, Christos Bergeles & Tom Vercauteren

SCIENTIFIC DATA

Tracking no: SDATA-25-04809A

ROBUST-MIPS: A Combined Skeletal Pose and Instance Segmentation Dataset for Laparoscopic Surgical Instruments

Authors: Zhe Han (King's College London), Charlie Budd (Kings College London), Gongyu Zhang (King's College London), Huanyu Tian (King's College London), Christos Bergeles (King's College London (KCL)), and Tom Vercauteren (Kings College London)

Abstract:
Localisation of surgical tools constitutes a foundational building block for computer-assisted interventional technologies. Works in this field typically focus on training deep learning models to perform segmentation tasks. Performance of learning-based approaches is limited by the availability of diverse annotated data. We argue that skeletal pose annotations are a more efficient annotation approach for surgical tools, striking a balance between richness of semantic information and ease of annotation, thus allowing for accelerated growth of available annotated data. To encourage adoption of this annotation style, we present, ROBUST-MIPS, a combined tool pose and tool instance segmentation dataset derived from the existing ROBUST-MIS dataset. Our enriched dataset facilitates the joint study of these two annotation styles and allow head-to-head comparison on various downstream tasks. To demonstrate the adequacy of pose annotations for surgical tool localisation, we set up a simple benchmark using popular pose estimation methods and observe high-quality results. To ease adoption, together with the dataset, we release our benchmark models and custom tool pose annotation software.

Datasets:

Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
Synapse	Robust Medical Instrument Pose and Segmentation (ROBUST-MIPS)		https://doi.org/10.7303/syn64023381	

ROBUST-MIPS: A Combined Skeletal Pose and Instance Segmentation Dataset for Laparoscopic Surgical Instruments

Zhe Han[1], Charlie Budd[1,*], Gongyu Zhang[1], Huanyu Tian[1], Christos Bergeles[1], and Tom Vercauteren[1]

[1]King's College London, School of Biomedical Engineering & Imaging Sciences, London, SE1 7EU, UK
[*]corresponding author: charles.budd@kcl.ac.uk

ABSTRACT

Localisation of surgical tools constitutes a foundational building block for computer-assisted interventional technologies. Works in this field typically focus on training deep learning models to perform segmentation tasks. Performance of learning-based approaches is limited by the availability of diverse annotated data. We argue that skeletal pose annotations are a more efficient annotation approach for surgical tools, striking a balance between richness of semantic information and ease of annotation, thus allowing for accelerated growth of available annotated data. To encourage adoption of this annotation style, we present, ROBUST-MIPS, a combined tool pose and tool instance segmentation dataset derived from the existing ROBUST-MIS dataset. Our enriched dataset facilitates the joint study of these two annotation styles and allow head-to-head comparison on various downstream tasks. To demonstrate the adequacy of pose annotations for surgical tool localisation, we set up a simple benchmark using popular pose estimation methods and observe high-quality results. To ease adoption, together with the dataset, we release our benchmark models and custom tool pose annotation software.

Background & summary

The localisation of surgical tools in intraoperative endoscopic video is a key capability in computer-assisted intervention (CAI). It holds the potential to facilitate novel CAI features such as safety analysis[1] and automated endoscope control[2], whilst also demonstrating a level of surgical scene understanding which could build into more complex technologies. While localisation can take many forms, the majority of works in this field focus on semantic segmentation, whereby a class label is predicted for every pixel in the image[3,4]. Additionally, some works have incorporated instance segmentation[5], a technique that extends semantic segmentation by distinguishing between individual instances of the same object class. Annotations for semantic segmentation require the creation of complex polygons or curves that follow the contours of each semantic object, be it a tool or a tool-part[6–8]. While these annotations provide detailed semantic information, they require significant time to create. In general-purpose computer vision domains, bounding boxes are often used to provide semantic information with minimal annotation effort. However, in the context of endoscopic video, the elongated and articulated structure of surgical tools makes bounding boxes less informative. They indeed often cover large portions of the image and significantly overlap with each other, reducing usefulness for precise localisation. We argue that skeletal pose annotations, such as those used in the field of human pose estimation[9], strike a better balance between semantic information and ease of annotation. Furthermore, skeletal pose annotations offer the additional benefit of capturing structural information, since they can help localise the tip and the shaft area, as well as instance-related information, since they can effectively distinguish between different instances of tools based on their unique skeletal structures. Thereby, they provide richer and more precise insights compared to traditional bounding box annotations.

Peng et al.[10] explore an alternative representation using tool tip bounding boxes combined with a line segment pointing along the tip. Backer et al.[11] propose a method using vector annotation to create detailed wireframes for surgical instruments. Instead of defining lines, different keypoints along the instruments are marked, allowing for precise representation of the instrument's structure and interactions within the surgical scene. Du et al.[12] provide tool pose annotations for 1,155 images from RMIT[13] and 1,850 from EndoVis[14]. These two datasets are limited due to their small size and significant redundancy, due to being tightly sampled from the source videos. In parallel to our efforts, Ghanekar et al.[15] proposed a multi-frame, context-driven model for video-based tracking of surgical tool keypoints. Their approach segments keypoint regions across consecutive frames using optical flow and monocular depth as auxiliary cues, followed by centroid estimation for localisation. This method demonstrated accurate tracking performance across challenging datasets such as EndoVis 2015[12] and JIGSAWS[16], highlighting the benefits of

temporal context in surgical tool-tip estimation. Wu et al.[17] introduced SurgPose to support more generalisable pose estimation in robotic surgery. The dataset comprises over 120k annotated instances across six types of da Vinci instruments, each with seven semantic keypoints, collected using a novel UV-based labelling method. It also includes stereo image pairs, kinematic data, and joint states, enabling both 2D and 3D pose estimation. SurgPose provides a strong foundation for vision-based pose estimation in robotic surgery, although it is currently limited to ex vivo environments and does not yet include complex scenes with mutual occlusions or inter-instrument interactions. The PhaKIR (`https://phakir.re-mic.de`) , reported in [18], advanced research in surgical instrument analysis by introducing a real-world multi-centre dataset with joint annotations for instance segmentation, keypoint estimation, and surgical phase recognition. However, the keypoint estimation task proved particularly challenging, with only two teams submitting and both achieving limited performance, largely due to instrument variation, occlusion, and class imbalance. This further motivated the design of ROBUST-MIPS to better support research on robust and generalisable tool pose estimation.

We aim to better establish the subfield of tool pose estimation by releasing ROBUST-MIPS (Medical Instrument Pose and Segmentation), a larger and more varied tool pose dataset, providing pose annotations for all 10,040 images of the ROBUST-MIS (Medical Instrument Segmentation) dataset[8,19]. As each frame also has tool instance segmentations, we hope that this could be used to investigate the strengths and weaknesses of these two annotation approaches as well as the interplay between these two tasks.

Methods

We outline the methodology for creating the skeletal pose representation of various surgical instruments. First, we present the data sources, as well as the origin and application context of the raw images used for annotation. Next, we detail the labelling protocol, including the number and type of keypoints defined for each instrument, along with their types and visibility labels. We then discuss the software tool and techniques employed to address annotation challenges. Finally, we provide a comprehensive description of the annotation process, covering the step-by-step procedures and information about the annotators involved in creating the ROBUST-MIPS dataset.

Data Sources

The ROBUST-MIPS dataset is derived from the ROBUST-MIS dataset, which was created for the ROBUST-MIS 2019 challenge[8]. This challenge aimed to benchmark algorithms for instrument segmentation and detection in minimally invasive surgery. It comprises 10,040 laparoscopic frames extracted from 30 colorectal surgical procedures, including 10 rectal resections, 10 proctocolectomies, and 10 sigmoid resections, all performed at Heidelberg University Hospital. All data were acquired using a Karl Storz laparoscopic camera system and downsampled to 960x540 pixels for computational efficiency. Ethical and legal considerations were addressed by fully anonymising all images, making the dataset suitable for public release without additional ethics approval[8].

Frames were sampled at 1 frame per second, with additional frames extracted during surgical phrase transitions to ensure sufficient coverage of varying surgical contexts. Each frame was provided with a pixel-wise instance segmentation mask indicating surgical instruments. ROBUST-MIS intentionally includes challenging imaging conditions typical of real-world surgical scenarios, such as bleeding, smoke, illumination changes, overlapping instruments, and partially visible tools.

The dataset was structured to support multiple tasks, including binary segmentation, multi-instance detection, and multi-instance segmentation, and was divided into training and testing sets to facilitate evaluation under increasing domain shifts. The organisation of files and folders in the dataset is illustrated in Figure 1. The dataset includes a detailed split strategy for accessing algorithm performance. In particular, the testing set was split into three stages reflecting escalating domain gaps: Stage1 used data from the same patients as training, Stage2 from new patients but the same surgery type, and Stage3 from a different surgery type.

The Labelling Protocol

To support the development of surgical instrument skeletal pose estimation, we have enriched the ROBUST-MIS dataset with additional annotations specifically designed for our target task. To ensure consistent and generalisable annotations, we established a rigorous labelling protocol covering three main topics: keypoint selection, annotation scope, and annotation guidelines.

Keypoints Selection

In the skeletal pose representation method, keypoint selection is guided by the type and characteristics of surgical tools. In our ROBUST-MIPS dataset, the keypoints are categorised into 4 main types:

- `EntryPoint`: In minimally invasive surgery, images captured through an endoscope typically have a circular content area[21]. The intersection point between the surgical instrument shaft and the circular content area boundary is defined

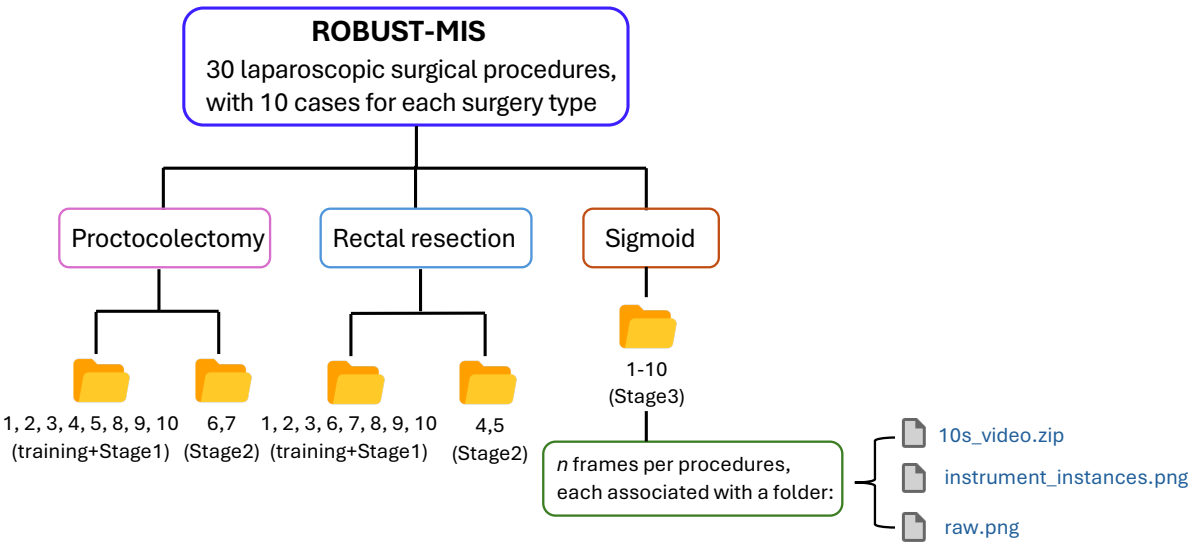


Figure 1. Overview of ROBUST-MIS data, the source data for the proposed ROBUST-MIPS. The directory structure is derived from the original ROBUST-MIS dataset[20]. While the original dataset provides a 10-second video snippet (250 frames) with the last raw frame with its instance segmentation mask, the proposed ROBUST-MIPS dataset extends this structure by incorporating skeletal pose annotations. The final extended directory structure of our contribution is detailed in Figure 6.

as the `EntryPoint`. These points are represented by red dots in Figure 2. Unlike structural landmarks (e.g., the `HingePoint`), the `EntryPoint` is not a fixed location on the instrument but varies dynamically as the tool moves in and out of the field of view (FoV). This definition is consistent with previous work in surgical pose estimation[12].

- `HingePoint`: For rigid surgical instruments, the intersection between the shaft and the metal or plastic tip is defined as the `HingePoint`, as shown in Figure 2(b). For articulated surgical instruments, the joint is considered the `HingePoint`, as shown in Figure 2(a). These points are indicated by green dots in Figure 2.
- `Tip1`/`Tip2`: The endpoints of all instruments can be labelled as tip points. For rigid instruments, there is only one endpoint, labelled as `Tip1`, as shown in Figure 2(b). For articulated instruments, there are two possible cases for endpoints: one endpoint, labelled as `Tip1`, or two endpoints, arbitrarily labelled as `Tip1` and `Tip2`. Taking graspers as a typical example, due to the structural symmetry of the jaws and their continuous axial rotation during surgery, defining a fixed *left* or *right* orientation is ambiguous. Even in cases where the two tips could be disambiguated (e.g. curved scissors), there is no universal convention across instruments to consistently order the tips. Furthermore, distinguishing subtly differing tips becomes unreliable under occlusion, smoke or rapid motion. Enforcing a semantic tip distinction in such cases would introduce significant label noise without adding much value to the geometric understanding. Consequently, `Tip1` and `Tip2` are defined as an unordered set in our dataset. Therefore, although `Tip1` and `Tip2` are distinguished by blue and yellow dots in Figure 2 for visualization, this assignment is interchangeable and permutation-invariant.

Figure 2 provides schematic examples of an articulated surgical tool and a rigid surgical tool, each illustrating the typical placement of these keypoints. For articulated instruments such as bipolar clamps, blunt graspers, and scissors, which have tips that can open and consist of two parts (shaft and tip), four keypoints are typically annotated: `EntryPoint`, `HingePoint`, `Tip1`, and `Tip2`. Rigid instruments like dissection hooks and probes also consist of a shaft and a tip, but their tips cannot open, resulting in three annotated keypoints: `EntryPoint`, `HingePoint`, and `Tip1`. Importantly, in our ROBUST-MIPS dataset, every surgical instrument is annotated using the same four keypoint categories: `EntryPoint`- `HingePoint`-`Tip1`-`Tip2`. However, the visibility status of each keypoint is crucial and is explicitly stored for each frame. The visibility labels are categorized into three states based on the visibility and inferability of the keypoint:

- `visible`: The keypoint is clearly visible in the image.
- `occluded`: The keypoint is not directly visible (e.g., covered by tissue or located slightly outside the FoV) but its position can be reliably inferred based on the instrument's mostly rigid geometry or symmetry.
- `missing`: The keypoint is completely out of view with insufficient cues for inference, or physically does not exist (e.g., the second tip of a rigid tool).

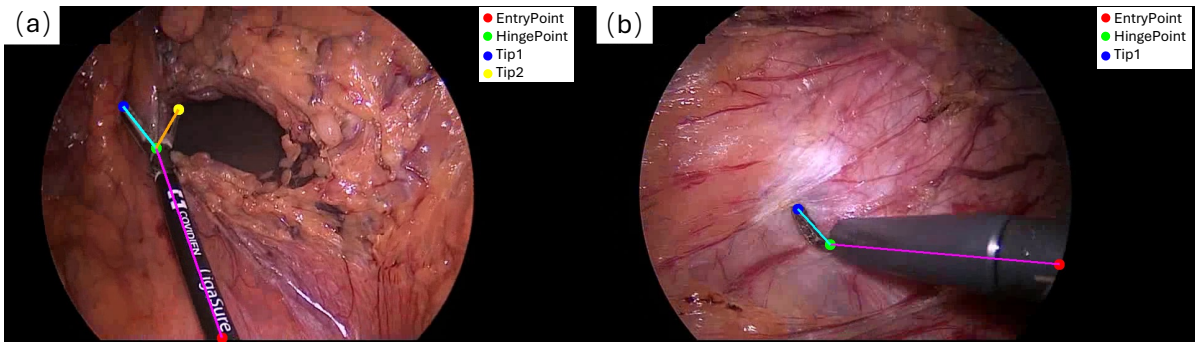


Figure 2. Examples of selecting keypoints for different types of surgical instruments. (a) Keypoints selected for an articulated surgical instrument. (b) Keypoints selected for a rigid surgical instrument.

While the definition of `visible` keypoints appears straightforward, the assignment of `missing` and `occluded` labels in specific scenarios can be practically challenging depending. Illustrative cases are shown in Figure 2 and Figure 3, detailed as follows:

- Scenarios for `missing`: This label is applied when a keypoint is physically absent or cannot confidently be inferred by the annotator. In terms of geometric constraints, since rigid instruments lack a second tip (cf. Figure 2 (b)) and closed articulated instruments have overlapping tips (cf. Figure 3 (b)), the fourth keypoint is marked as `missing` to maintain the data format. Regarding visual ambiguity, when the distal end is entirely hidden and visual cues are insufficient for inference by the annotator, the keypoints are defined as `missing`. This includes cases where the tip is entirely hidden (cf. Figure 3 (a)) or only the shaft is visible (cf. Figure 3 (c)).

- Scenarios for `occluded`: This label applies to keypoints that are physically present but visually occluded. For geometric inference within the image frame, if a tip is covered by tissue but inferable via symmetry (cf. Figure 3 (f)), or if a keypoint falls into the non-informative image region outside the circular FoV but remains within the rectangular image boundary (cf. Figure 3 (e)), they are labelled as `occluded` with valid positive coordinates. To maintain the connectivity of the annotation chain, the `EntryPoint` serves as the root node, it also be labelled as `occluded` in Figure 3 (e). For skeletal connectivity beyond the image boundary, we allow the annotator to label points outside of the image boundary. As illustrated in Figure 3 (d), the instrument shaft may indeed be inferred to be completely outside the frame, meaning both the `EntryPoint` and `HingePoint` are located outside the image boundary. To accommodate this, our annotation software provides a zoom-out function and a designated padding area, allowing these points to be annotated with out-of-bounds coordinates. While these points ensure structural completeness, they are effectively filtered out during training due to their out-of-bounds values.

Table 1 summarises the skeletal representations of surgical instruments, combining these variations in instrument types with differences in operational or visibility states.

Table 1. Overview of the skeletal representation of articulated and rigid surgical tools in different states, with the visualisation of each case shown in Figure 2 and Figure 3.

Tool types	States	Tool representation	Cases
Articulated	all keypoints visible	4 points and 3 lines	Figure 2(a)
	Tips or HingePoint occluded		Figure 3(d,e,f)
	one tip missing/closed	3 points and 2 lines	Figure 3(a,b)
	only shaft in the FoV	2 points and 1 line	Figure 3(c)
Rigid	all keypoints visible)	3 points and 2 lines	Figure 2(b)
	only shaft in the FoV	2 points and 1 line	similar with the Figure 3(c)

Annotation Scope

The annotation scope defines the entities to be annotated and excluded to ensure a focused and high-quality dataset suitable for surgical instrument pose estimation. While the ROBUST-MIS dataset provides instance segmentation masks for various

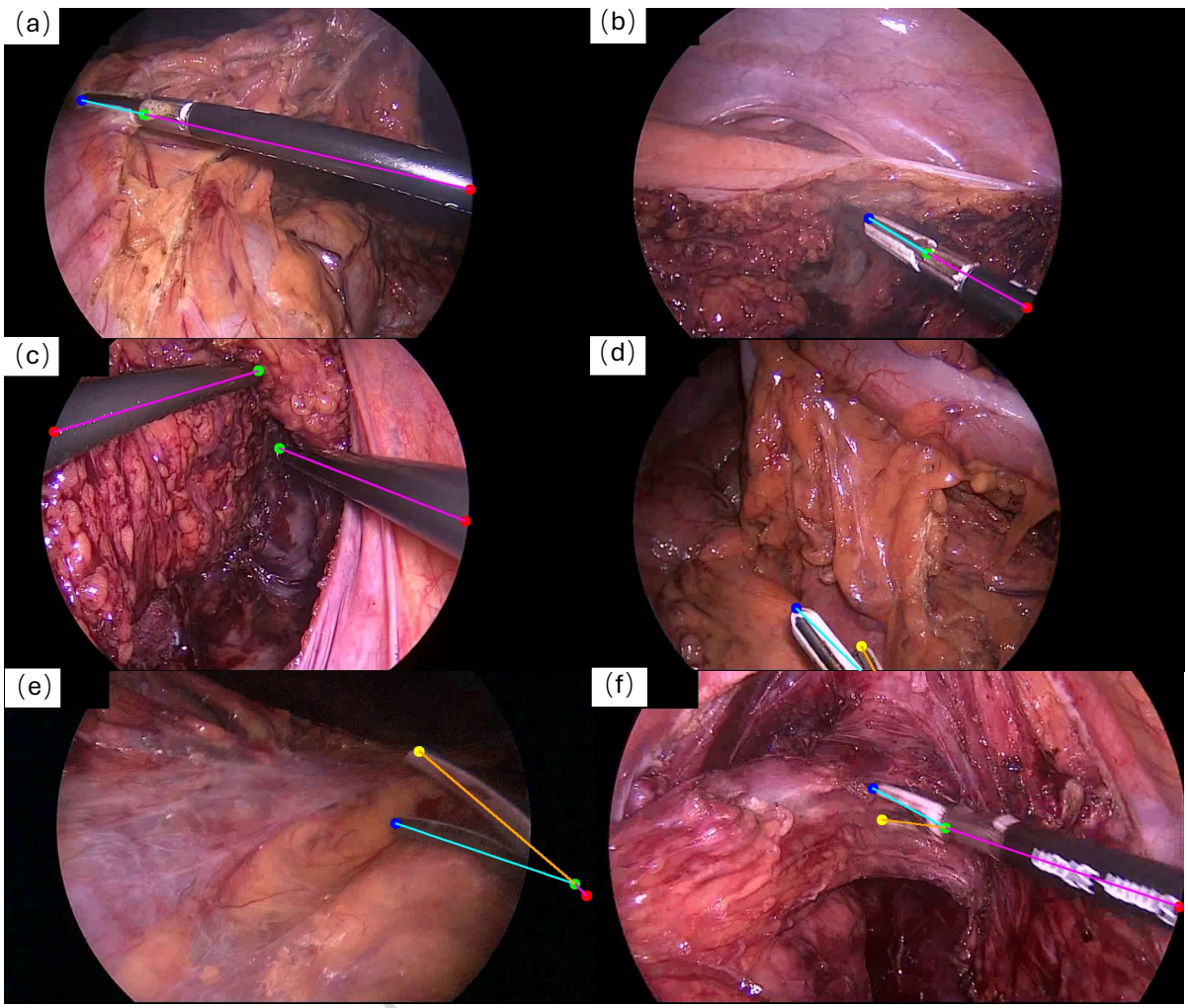


Figure 3. Examples of selecting valid keypoints in different visibility. (a) Selection of keypoints for partially occluded articulated surgical tools, where one tip point is considered to be in an unpredictable missing state. (b) Selection of keypoints for articulated surgical tool in a closed state, where the two tips are considered to be at the same position. The second tip is labelled as `missing`. (c) Selection of keypoints for surgical tools with only the shaft visible in the FoV, resulting in `missing` labels for both tips. (d) Selection of keypoints when the instrument shaft extends beyond the image boundary. To maintain skeletal connectivity, both the `HingePoint` and `EntryPoint` are annotated in the padding area with out-of-bounds coordinates, despite being strictly invisible. (e) Selection of keypoints where the `HingePoint` is masked by the circular FoV but remains within the image frame. The point is geometrically predicted from the tips and arm structure, possessing valid positive coordinates. In this case, the `HingePoint` and `EntryPoint` are both labelled as occluded. (f) Selection of keypoints where one of the tips is inferred based on the instrument's structural characteristics, such predicted keypoints are annotated as occluded.

surgical tools and has been widely used for surgical scene analysis, adapting it for pose estimation revealed certain limitations in its original annotation protocol, particularly regarding the labelling of trocars and cannulas/ports.

In the context of segmentation, trocar cannulas are tubular devices that serve as ports during laparoscopic surgery and are considered visible structures that must be separately identified from surrounding tissues and surgical instruments. Accordingly, the ROBUST-MIS dataset annotates both camera trocar cannulas, which hold the endoscope, and tool trocar cannulas, which serve as entry points for surgical instruments, as individual instances.

However, for pose estimation tasks, this level of detail introduces challenges:

- Camera trocar cannulas: These are static structures fixed to the patient or surgical robot, contributing no dynamic motion or orientation information relevant for instrument pose estimation.

- Tool trocar cannulas: Although physically connected to the instruments, tool trocar cannulas merely represent a fixed entry point into the surgical field and do not reflect the dynamic geometry or movement of the instrument itself.

Including trocar cannulas in pose annotations can therefore introduce unnecessary noise and redundancy, as the functional pose of surgical instruments is defined by the shaft and tip beyond the trocar cannula rather than the trocar cannula itself. When an instrument extends from a trocar cannula, we define the distal end of the trocar cannula as the `EntryPoint` for pose annotation, as shown in Figure 4 (a), while other annotation principles remain unchanged. For consistency, in the instance segmentation annotations of ROBUST-MIPS, we removed the masks corresponding to both camera trocar cannulas and tool trocar cannulas that were present in the original ROBUST-MIS dataset. Taking the tool trocar cannulas as an example, Figure 4 (b) and (c) illustrate this process by comparing the original instance label with our refined mask. It should also be noted that frames containing only camera trocar cannulas do not possess corresponding skeletal annotations for surgical instruments. Consequently, removing the camera trocar masks from these images results in completely empty (black) segmentation labels.

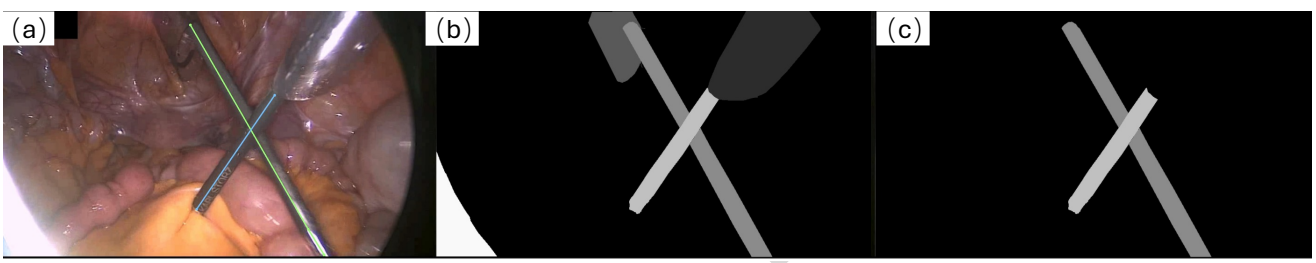


Figure 4. Handling of tool trocar cannulas in ROBUST-MIPS annotations. (a) Pose annotation example in the presence of a tool trocar cannula. The distal end of the cannula is identified as the `EntryPoint`, while the instrument shaft extends to the `HingePoint`. (b) The original segmentation annotation in ROBUST-MIS[20], where the trocar cannula is labelled as a distinct instance. (c) The refined segmentation annotation in ROBUST-MIPS dataset[22], where the trocar cannula mask is removed to focus solely on the surgical instrument.

Annotation Guidelines

ROBUST-MIS intentionally includes challenging imaging conditions typical of real-world surgical scenarios, such as bleeding, smoke, illumination changes, overlapping instruments, and partially visible tools. To address these challenging cases, we provided specialised instructions and examples for annotators. These included guidance on reviewing the corresponding 10-second video clips for clarification, referring to already annotated segmentation masks from ROBUST-MIS when uncertain, annotating as much of the instrument as possible, and continuing annotations even under low visibility conditions. In some cases where it is not possible to distinguish between the shaft and the tip, or where no clear boundary exists between these parts, only the `EntryPoint` and `HingePoint` keypoints are annotated. Furthermore, additional factors such as motion blur, reflections, lens dirtiness, and the presence of fluids were also considered during annotation, ensuring robust and consistent labelling across diverse surgical scenes.

Annotation Software

To support the annotation process for our ROBUST-MIPS dataset, we designed open-source annotation software specifically for manual surgical instrument pose labelling, with its source code available on GitHub(`https://github.com/cai4cai/tool-pose-annotation-gui`). The software provides a graphical interface that enables efficient image browsing and intuitive keypoint annotation. Specifically, semantic abbreviations (e.g., 'E' for `EntryPoint`, 'T1' for `Tip1`) are displayed next to each point to aid in identifying instrument parts. Users can zoom out with the mouse scroll wheel. This function is particularly useful for placing occluded keypoints located outside the visible image area. Annotation begins with a left click, which either starts a new pose or adds a keypoint to the current one. Right click is used to annotate occluded keypoints by placing an estimated position. Middle click completes the current pose annotation. Clicking on the edge of an existing skeleton allows users to insert a visible/occluded transition point, and the remaining keypoint tags are automatically updated. Our custom annotation software also enables efficient mask removal to ensuring that the instance segmentation masks are better suited for surgical instrument pose estimation tasks. This software ensures efficient and consistent annotations across various surgical instrument types and visibility conditions, as described above.

Data Description

The keypoint information obtained through the annotation software is stored as a JSON file using the schema shown in Figure 5, with one such JSON file generated for each image. In this schema, `nodes` contains the coordinates of the keypoints, while `tags`

records their visibility status. The `edges` field specifies the connections between pairs of keypoints. The `transitions` field represents intermediate points located between visible keypoints and non-visible keypoints (either occluded or missing). As illustrated in Figure 5(a), if a tip point is visible but its connected hinge point is not, there must be a segment between them that is visible in the image. In this case, the farthest visible point along the arm from the tip is recorded as a `transition` point.

(a)


```
[
  {
    "nodes": [
      [247.9, 533.9],
      [208.0, 408.1],
      [149.1, 244.2],
      null
    ],
    "tags": ["visible","visible","visible","missing"],
    "edges": [[0, 1],[1, 2],[1, 3]],
    "transitions": [[],[],[]]
  },
  {
    "nodes": [
      [390.8, 567.9],
      [381.3, 555.0],
      [337.5, 499.6],
      null
    ],
    "tags": ["occluded","occluded","visible","missing"],
    "edges": [[0, 1],[1, 2],[1, 3]],
    "transitions": [[],[[366.6, 536.5]],[]]
  }
]
```


(b)


```
{
  "categories": [
    {
      "supercategory": "SurgicalTool",
      "id": 1,
      "name": "SurgicalTool",
      "keypoints": ["entry", "hinge", "tip1", "tip2"],
      "skeleton": [[0, 1], [1, 2], [1, 3]]
    }
  ],
  "images": [
    {
      "file_name": "file_dir/Stage2_Proctocolectomy_6_1500.png",
      "height": 540,
      "width": 960,
      "id": 0
    }
    ...
  ],
  "annotations": [
    {
      "category_id": 1,
      "image_id": 0,
      "id": 0,
      "bbox": [22, 52, 302, 295],
      "area": 89114.5,
      "keypoints": [42.5, 327.5, 2, 159.2, 219.2, 2,
                    106.7, 72.5, 2, 304.2, 123.3, 2],
      "num_keypoints": 4
    }
    ...
  ]
}
```


Figure 5. (a) Example of JSON annotation file from the custom annotation software. (b) Example of an annotation converted to the Microsoft COCO schema[23] which allows for broad compatibility with human pose learning framework.

Annotation Procedures

This section provides an overview of the workflow adopted to construct the ROBUST-MIPS dataset, along with insights into the human annotation process. The description is organised into two main topics: The step-by-step procedure and the role of annotators.

Step-by-Step Procedure

1. Base dataset selection and preparation: The ROBUST-MIPS dataset was developed based on the existing open-access dataset ROBUST-MIS, which provided raw laparoscopic video data without instance-level keypoint annotations.

2. Development of the labelling protocol: Prior to annotation, we conducted extensive discussions to design a comprehensive labelling protocol. This protocol specified definitions for keypoint selection and rules for handling challenging scenarios, such as partial occlusion, poor lighting conditions, and camera trocar cannulas.

3. Manual annotation of keypoints: All keypoint annotations were performed entirely manually using a custom annotation software. Beyond spatial coordinates, the visibility status of each keypoint (visible, occluded, missing) was incorporated into the annotation process. Annotators labelled each frame individually, specifying the positions of predefined keypoints for visible instruments. In cases where certain keypoints were not visible, their status was recorded as either occluded or missing, following the defined protocol. This additional information is crucial for downstream analysis and model training.

4. Final quality control: Upon completion of annotations, thorough quality control procedures were carried out. Each annotated frame was reviewed manually to verify correctness and consistency with the labelling protocol. Additionally, the entire dataset underwent a second round of review by a different annotator to ensure accuracy and to resolve any potential discrepancies.

Annotators

The annotation process for creating the ROBUST-MIPS dataset involved a primary annotator responsible for the majority of annotation tasks and quality control. A secondary annotator, with greater experience, assisted in annotation and quality control. Both annotators were supported by an expert team to ensure accuracy and consistency in the annotations.

Data Records

The ROBUST-MIPS dataset[22] is accessible via the public repository on Synapse. The imaging data utilized in this work derive from the publicly available ROBUST-MIS dataset[20], which is also hosted on Synapse. Our derived dataset is distributed as a compressed archive file, `ROBUST-MIPS.zip`, which contains three components for each frame: the raw endoscopic image (`raw.png`), the corresponding instance segmentation mask (`instrument_instances.png`), and the keypoint annotation file (`raw.json`), as illustrated in Figure 6.

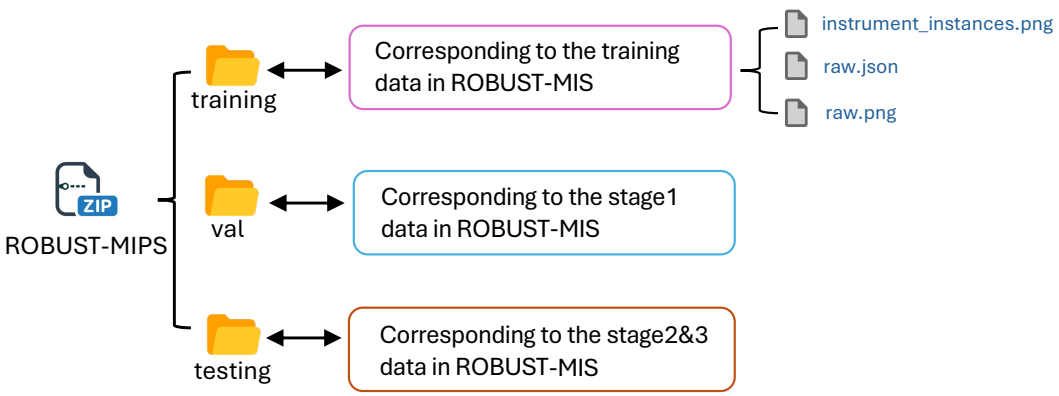


Figure 6. Directory structure for ROBUST-MIPS dataset. Each split contains subdirectories following a `Surgery_type/Procedure_ID/Frame_ID` structure.

To enhance the generalisability of the dataset and to facilitate subsequent performance evaluations on various popular pose estimation models, as discussed in the *Models* Section, we processed the annotation information corresponding to each image and consolidated it into a JSON format similar to that used in the Microsoft COCO (Common objects in context) dataset[23], as illustrated in Figure 5(b). In our dataset, all instruments are grouped into a single category, as shown in the `categories` object. The `images` object contains information about the file paths of each image, along with their resolutions and unique identifiers. The `annotations` object comprises a list in which each entry corresponds to the keypoint annotations for a specific frame. The `image_id` corresponds to the id in the `images` object, while the `id` refers to the identifier of the current surgical tool. Keypoint annotations are expressed as (x, y, v), where x and y correspond to the horizontal and vertical coordinates of the keypoint, with the origin of the coordinate system at the top left of the image. v indicates the visibility attribute of the keypoint. For each annotated keypoint, the value of the visibility property indicates if a keypoint is annotated and visible ($v = 2$), annotated and occluded ($v = 1$), or not annotated because it is not located inside the frame or in case it is not possible to estimate its position accurately ($v = 0$). The `num_keypoints` represents the number of keypoints with a v-value greater than 0.

Additionally, the JSON file also includes the bounding box information for each instrument, which is calculated based on the coordinates of the keypoints and is denoted as $[x_{min}, y_{min}, w, h]$. The `area` object is computed as the square of the diagonal of the bounding box.

Bounding Box Generation

To enable our dataset to support training based on both the top-down and bottom-up paradigms for pose estimation tasks[24], the JSON files also include the bounding boxes calculated from the 2D keypoints for each surgical tool. The top-left corner of each bounding box is defined by the minimum x and y coordinates, denoted as (x_{min}, y_{min}). The width w and height h of the bounding box are calculated as the differences between the maximum and minimum x coordinates, $x_{max} - x_{min}$, and the maximum and minimum y coordinates, $y_{max} - y_{min}$, respectively. Thus, the bounding box for each tool can be accurately represented as $[x_{min}, y_{min}, w, h]$. While the method of generating bounding boxes based on skeletal information effectively represents the pose of surgical instruments in most cases, as shown in Figure 7(a), it performs poorly when the surgical tool is in a horizontal or vertical position within the FoV, as illustrated in Figure 7(c,e). This is because the vertical or horizontal coordinates of the keypoints used for the calculation are too close to each other, resulting in overly narrow bounding boxes. Additionally, for some tools with curved shapes, the bounding boxes calculated solely based on keypoint coordinates may not adequately represent the entire tool, as shown in Figure 7(g).

To address this issue and improve the accuracy of the bounding boxes generated from the 2D keypoint annotations, a margin of 20 pixels is added to the calculated boundaries on all sides. If the expanded bounding box exceeds the image boundaries, the image boundaries are used as the limit.

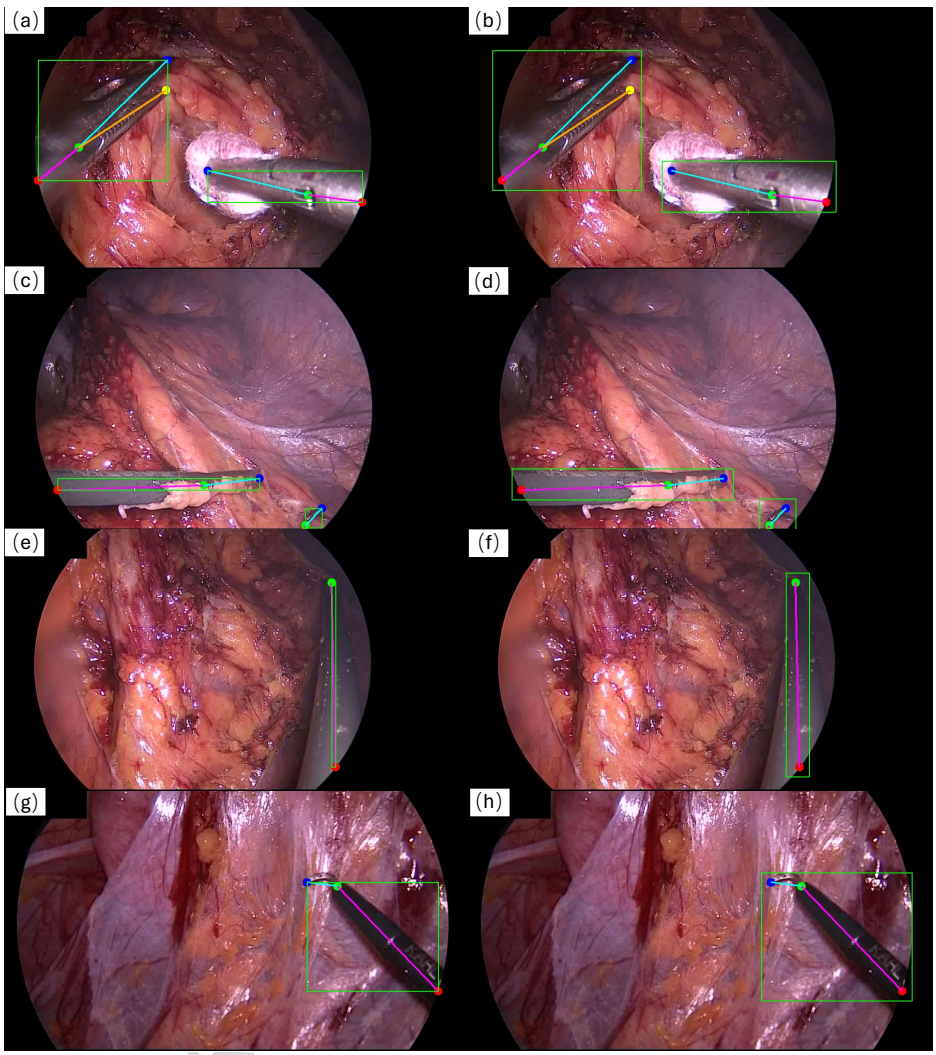


Figure 7. Examples of bounding box generated from 2D keypoints. (a) Bounding boxes that can generally represent the surgical tool. (b) Result after adding a margin to the bounding box in (a). (c) Bounding box result for a surgical tool in a horizontal position within the FoV. (d) Result after adding a margin to the bounding box in (c). (e) Bounding box result for a surgical tool in a vertical position within the FoV. (f) Result after adding a margin to the bounding box in (e). (g) Bounding box result for a surgical tool with a curved shape. (h) Result after adding a margin to the bounding box in (g).

Dataset Split

In the ROBUST-MIS Challenge 2019, the dataset was divided as shown in Table 2 to evaluate the generalisability and performance of algorithms. The ROBUST-MIPS dataset has similarly been structured but has been partitioned into training, validation, and testing sets, as shown in Table 3. It is important to note that images in Stage1 originate from the same patients as those in the training set; therefore, Stage1 data is used as the validation set in the ROBUST-MIPS dataset. Data from Stage2 and Stage3 have been allocated to the testing set to enable a comprehensive evaluation of the model generalisation ability.

Technical Validation

This section presents a comprehensive overview and technical validation of our ROBUST-MIPS dataset, including showcases and performance evaluation. The performance evaluation results demonstrate its reliability and effectiveness for use in research and development of pose estimation models.

Table 2. Case distribution of the data with frames per stage and surgery of the ROBUST-MIS dataset. Empty frames (denoted as ef in the table) were classed as the % of frames in which an instrument did not appear.

Procedure	Training	Testing		
		Stage 1	Stage 2	Stage 3
Proctocolectomy	2,943(2% ef.)	325 (11% ef.)	225 (11% ef.)	0
Rectal resection	3,040 (20% ef.)	338 (20% ef.)	289 (15% ef.)	0
Sigmoid resection	0	0	0	2880 (23% ef.)
Total	5983 (17% ef.)	663 (15% ef.)	514 (13% ef.)	2880 (23% ef.)

Table 3. Case distribution of the data with frames per stage and surgery of ROBUST-MIPS dataset. The training and validation data come from the same group of patients undergoing two types of surgeries, while the testing set includes data from different patients undergoing the same two surgery types, as well as a third surgery type not present in the training process.

Procedure	Training	Validation	Testing
Proctocolectomy	2,943	325	225
Rectal resection	3,040	338	289
Sigmoid resection	0	0	2880
Total	5983	663	3394

Training and Evaluation of Baseline Models

Models

To validate the usability of the ROBUST-MIPS dataset, three baseline pose estimation models, RTMPose[25], SimpleBaseLine[26], and ViTPose[27] were trained. To establish benchmark performance metrics for future researchers to use as a comparison, we chose to utilise a range of baselines in pose estimation tasks. Although these models were originally designed for human pose estimation, we employed them to explore their generalisability in surgical tool pose estimation. Each keypoint of a surgical tool can be regarded as a joint in human pose estimation tasks. However, when annotating the human skeleton, symmetric points such as the left and right shoulders, or the left and right elbow joints, have distinct physical meanings, and consistency in annotation must be maintained across different individuals. In contrast, for symmetric surgical tools like scissors or forceps, the two tip points, `Tip1` and `Tip2`, do not require a specific order in the annotation. When using the aforementioned models for surgical tool pose estimation, the equivalence between the tips of the surgical tool has not been addressed.

The models were trained using the training and validation splits of the ROBUST-MIPS dataset, with hyperparameters detailed in Table 4. All training was conducted on an NVIDIA A100 32G GPU. The MMPose open-source tool[28] was utilised for both training and evaluation, as it includes a comprehensive set of relevant models and evaluation metrics. For data augmentation, the MMPose framework implements standard techniques such as cropping, flipping, color distortion, rotation, and scaling.

Regarding the optimization objectives, we utilized the specific loss implementations provided by the MMPose framework. RTMPose was trained using KLDiscretLoss, which computes the KL divergence between predicted and ground-truth distributions on discretized coordinates[29]. In contrast, SimpleBaseLine and ViTPose utilized KeypointMSELoss. Unlike standard MSE or KL loss functions, these framework-specific implementations inherently incorporate a target weight masking mechanism. This ensures that keypoints annotated as `missing` are effectively excluded from backpropagation, thereby preventing invalid gradients from affecting the training process.

Table 4. Parameters of the models.

optimiser	AdamW
base learning rate	0.0005
learning rate schedule	LinearLR
batch size	32(train) 16(val)
warm-up iterations	500
weight decay	0.01
training epochs	600

Recommended Metric

The COCO Object Keypoint Similarity (COCO OKS) metric[23] is designed to provide a quantitative assessment of the similarity between predicted keypoints and ground truth keypoints, taking into account the scale of the object and the relative importance of different keypoints:

$$OKS = \sum_i [\exp(-\frac{d_i^2}{2s^2\kappa_i^2})\delta(v_i > 0)] / \sum_i [\delta(v_i > 0)] \tag{1}$$

where d_i is the Euclidean distance between the predicted keypoint and the ground truth keypoint. κ_i is a per-keypoint constant that controls falloff, which helps in normalising the effect of different keypoints. v_i is the visibility flags of the ground truth (the predicted visibility tags are not used). Predicted keypoints that are not labelled ($v_i = 0$) do not affect the OKS. In (1), s is the scale of the object.

As discussed above, the two tip points, `Tip1` and `Tip2`, do not require a specific order in the annotation. In the COCO OKS, the equivalence between the tips of the surgical tool is not addressed. We propose a simple modification of the metric where a version of the ground truth pose is constructed by swapping the order of the two tips. We evaluate the OKS of a prediction against both the initial and tip-swapped ground truth and report the best value. This achieves the same outcome as including a bipartite matching step as suggested in the PhaKIR challenge[18].

In standard COCO OKS, s is defined as the square root of the object bounding box area ($s = \sqrt{wh}$). However, as shown in Figure 7, surgical tools are typically slender, elongated structures with high aspect ratios (length $\gg$ diameter). For such objects, the area of an axis-aligned bounding box is highly sensitive to 2D rotation. It collapses to near-zero when the tool is axis-aligned (horizontal or vertical) but expands significantly when rotated diagonally. Using the standard definition would thus result in an inconsistently strict metric, penalising axis-aligned poses disproportionately due to this area collapse. To address this, we redefine s based on the arithmetic mean of the squared dimensions (scaled diagonal):

$$s = \sqrt{\frac{w^2 + h^2}{2}} \tag{2}$$

Based on the inequality of arithmetic and geometric means, this formulation ensures $s^2 \geq wh$, providing a robust scale factor that is dominated by the tool's length (diagonal) rather than its projection width. This guarantees rotation invariance and prevents the evaluation scale from becoming excessively small in axis-aligned states. For square-like objects (where $w \approx h$), the formula $\frac{w^2+h^2}{2}$ mathematically reduces to the standard area ($w \cdot h$). This ensures that our metric remains consistent with the original definition of s for non-slender shapes.

In the COCO human pose dataset, for each keypoint type i, a standard deviation σ_i is defined to capture the expected human annotation uncertainty. These σ_i are also used to score the quality of automatically predicted keypoints against manual annotations. This value essentially reflects the difficulty of consistently labelling a specific anatomical landmark on the human body. For instance, distinct and well-localised landmarks like eyes are assigned a low scale-normalized σ_i (≈ 0.026), meaning that predictions must be very precise relative to the object size to score highly. Conversely, structurally ambiguous points or those covered by soft tissue, such as hips, are assigned a higher σ_i (≈ 0.107) which allows for a larger margin of error[23]. To make COCO OKS a perceptually meaningful metric, the keypoint-specific falloff constants are typically set as $\kappa_i = 2\sigma_i$.

In our work, we lack the large-scale repeated annotations required to empirically calculate σ_i for each specific instrument keypoint type. However, surgical tools present significant challenges, including diverse structural variations, tissue occlusion, and a lack of distinct surface textures (e.g., smooth metallic shafts), which inherently increase annotation variance. Therefore, we adopted a conservative strategy by relying on the maximum standard deviation σ_i used in the COCO human pose dataset. We thus set $\sigma_i = 0.107$ (corresponding to the 'hips' category) for all keypoints in our surgical tool annotation task. This choice acknowledges the geometric ambiguity of surgical tools and ensures that the evaluation metric remains fair and robust even under challenging visual conditions.

We utilize the official COCO implementation (`https://github.com/cocodataset/cocoapi/blob/master/PythonAPI/pycocotools/cocoeval.py`) for calculating the Average Precision (AP) and Average Recall (AR) metrics in keypoint detection. AP measures the precision of the model in detecting keypoints, and AR evaluates the model ability to recall objects over a range of thresholds. $\mathbf{AP}_k$ represents the average precision at a given OKS threshold k, while $\mathbf{AR}_k$ denotes the average recall at OKS threshold k. The default **AP(OKS)** and **AR(OKS)** are generally averaged over multiple OKS values, specifically from 0.50 to 0.95 with increments of 0.05.

Additionally, since Top-Down methods rely on a two-stage pose estimation model, the first stage involves bounding box detection. The evaluation of this stage aligns with the evaluation methods used in object detection tasks, utilising the COCO Intersection over Union (IoU) metric. IoU measures the overlap between the predicted bounding box and the ground truth bounding box, calculated as the ratio of the area of intersection to the area of union. Likewise, IoU plays a key role in calculating

AP and AR. $\mathbf{AP_{IoU=}}_k$ represents the average precision at a given IoU threshold k, and **AP(IoU)** is the COCO-style metric, averaged over IoU thresholds from 0.5 to 0.95 in increments of 0.05. $\mathbf{AR_{IoU=}}_k$ denotes the average recall at IoU threshold k, and **AR (IoU)** follows the same COCO-style averaging process.

Performance Evaluation

The performance results are presented in Table 5, demonstrating the effectiveness of the dataset in training robust pose estimation models. The models were evaluated on the testing set that was not previously encountered during training or validation to assess their generalisation capability. A qualitative comparison between the model predictions on the testing set and the corresponding ground truth can be seen in Figure 8.

Table 5. Results of various algorithms for surgical tool pose estimation on the ROBUST-MIPS testing set. SBL stands for SimpleBaseLine[26].

Model	Backbone	Resolution	Robust-MIP testing					
			AP	$\mathbf{AP_{OKS=0.5}}$	$\mathbf{AP_{OKS=0.75}}$	**AR**	$\mathbf{AR_{OKS=0.5}}$	$\mathbf{AR_{OKS=0.75}}$
SBL	ResNet152	256x192	0.694	0.819	0.704	0.732	0.834	0.739
SBL	ResNet152	384x288	0.684	0.807	0.694	0.730	0.830	0.740
RTMPose	CSPNext-m	256x192	0.705	0.820	0.716	0.740	0.839	0.748
RTMPose	CSPNext-l	256x192	0.712	0.827	0.722	0.750	0.845	0.758
ViTPose-B	ViT-B	256x192	0.735	0.832	0.750	0.768	0.847	0.778
ViTPose-L	ViT-L	256x192	**0.754**	**0.842**	**0.771**	**0.784**	**0.855**	**0.796**

Limitations and Possible Improvements

The proposed dataset in this paper has a few limitations that must be taken into consideration. One of the main limitations is that not all surgical tools can be accurately represented using this scheme. For example, curved instruments like scissors or hooks present challenges. In the case of scissors, the shaft is not straight, so the line segment connecting the keypoints does not accurately represent the actual shape of the tool. Similarly, for hooks, the line connecting the tip and the hinge point fails to capture the curvature of the hook. Another limitations is that all the surgical tools are categorised under a single class without more detailed classification labels, such as forceps, hooks, scissors, needle drivers and so on. While this may be sufficient for the task of surgical instrument pose estimation, it limits the generalisability of the dataset.

In addition, the baseline models employed in this study predict the endpoints of surgical instruments independently, and the annotation process does not enforce a consistent order of these points across samples. As a result, the predicted endpoints may occasionally be assigned in a different order from the reference annotation. While the invariance with respect to tip order is already considered in our proposed modified OKS, the phenomenon nonetheless also underscores a modelling limitation. Future architectures may benefit from explicitly encoding tip-level equivalence or ordering invariance during training, which could enhance both prediction stability and semantic consistency.

Usage Notes

ROBUST-MIPS is released under a Creative Commons Attribution-NonCommercial ShareAlike license (CC BY-NC-SA), which is required to align with the non-commercial restrictions applied to the source data.

Data Availability

The ROBUST-MIPS dataset generated and analyzed in this study is publicly available at https://doi.org/10.7303/syn64023381 [22]. The imaging data used to construct this dataset were obtained from the publicly available ROBUST-MIS dataset, accessible via https://doi.org/10.7303/syn18779624 [20].

Code Availability

The annotation software is made public at https://github.com/cai4cai/tool-pose-annotation-gui. We also release the code for benchmark training at https://github.com/cai4cai/ROBUST_MIPS_toolpose. It also contains scripts for converting the data to the COCO format.

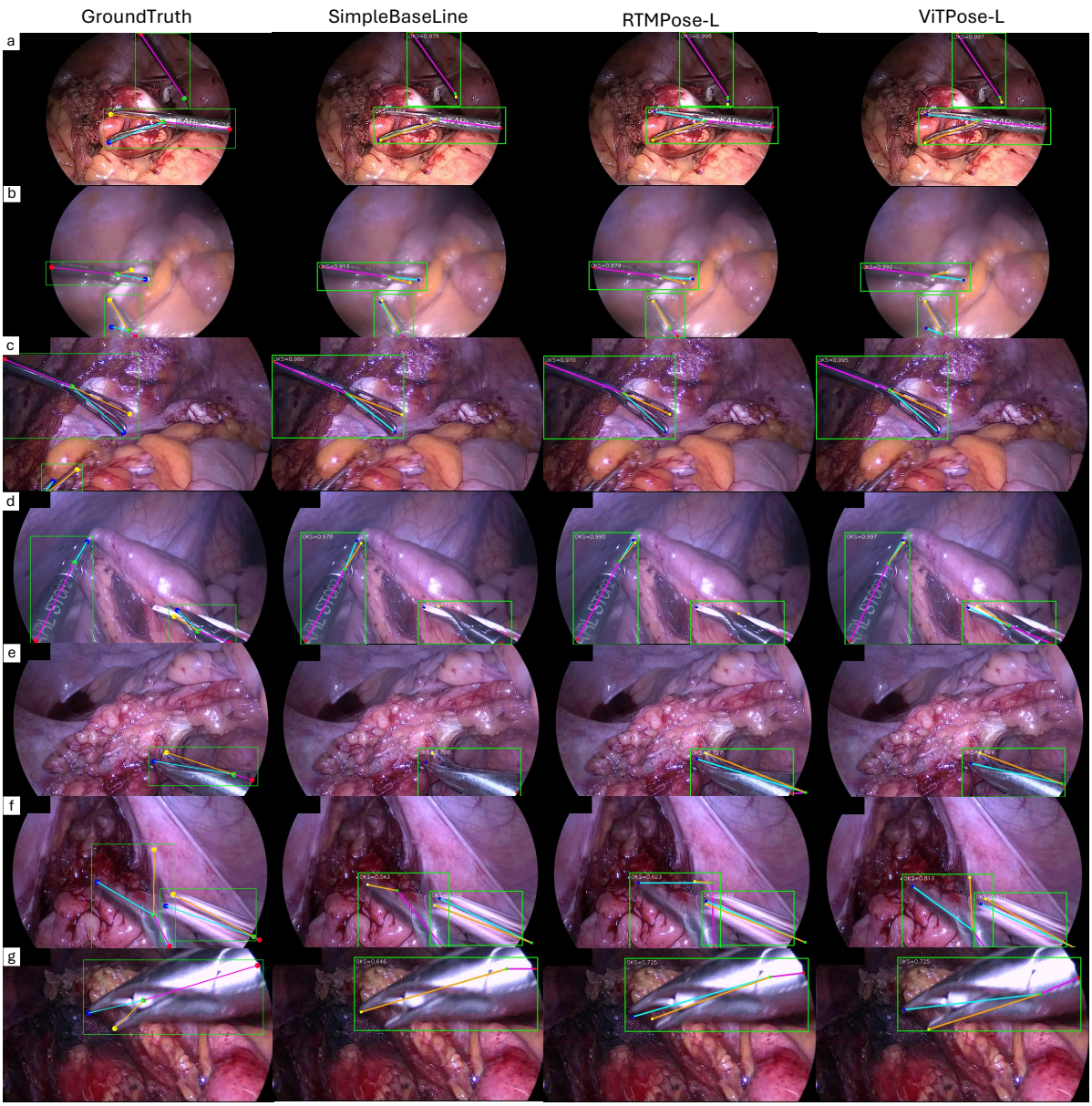


Figure 8. Visual representation of the performance of SimpleBaseLine, RTMPose, ViTPose and the corresponding ground truth annotations. OKS score values are indicated on each subfigure and these are better visualised by digital zoom in. Predicted keypoints that are labelled as missing ($v_i = 0$) do not affect the OKS.

References

1. Ríos, M. S. *et al.* Cholec80-CVS: An open dataset with an evaluation of Strasberg's critical view of safety for AI. *Sci. Data* **10**, 194 (2023).
2. Gruijthuijsen, C. *et al.* Robotic endoscope control via autonomous instrument tracking. *Front. Robotics AI* **9**, 832208 (2022).
3. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241 (Springer, 2015).
4. García-Peraza-Herrera, L. C. *et al.* ToolNet: Holistically-nested real-time segmentation of robotic surgical tools. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5717–5722 (2017).
5. Alabi, O. *et al.* CholecInstanceSeg: A tool instance segmentation dataset for laparoscopic surgery. *Sci. Data* **12**, 1–12 (2025).
6. Twinanda, A. P. *et al.* EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Med. Imaging* **36**, 86–97 (2016).
7. Hong, W.-Y. *et al.* CholecSeg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on Cholec80. *arXiv preprint arXiv:2012.12453* (2020).
8. Roß, T. *et al.* Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the ROBUST-MIS 2019 challenge. *Med. Image Analysis* **70**, 101920 (2021).
9. Cao, Z., Simon, T., Wei, S.-E. & Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299 (2017).
10. Peng, J., Chen, Q., Kang, L., Jie, H. & Han, Y. Autonomous recognition of multiple surgical instruments tips based on arrow OBB-YOLO network. *IEEE Transactions on Instrumentation Meas.* **71**, 1–13 (2022).
11. De Backer, P. *et al.* Multicentric exploration of tool annotation in robotic surgery: lessons learned when starting a surgical artificial intelligence project. *Surg. Endosc.* **36**, 8533–8548 (2022).
12. Du, X. *et al.* Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE Transactions on Med. Imaging* **37**, 1276–1287 (2018).
13. Sznitman, R. *et al.* Data-driven visual tracking in retinal microsurgery. In Ayache, N., Delingette, H., Golland, P. & Mori, K. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012*, 568–575 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
14. Reiter, A., Allen, P. K. & Zhao, T. Feature classification for tracking articulated surgical tools. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part II 15*, 592–600 (Springer, 2012).
15. Ghanekar, B., Johnson, L. R., Laughlin, J. L., O'Malley, M. K. & Veeraraghavan, A. Video-based surgical tool-tip and keypoint tracking using multi-frame context-driven deep learning models. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5 (IEEE, 2025).
16. Gao, Y. *et al.* Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, vol. 3, 3 (2014).
17. Wu, Z. *et al.* Surgpose: a dataset for articulated robotic surgical tool pose estimation and tracking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 10507–10514 (2025).
18. Rueckert, T. *et al.* Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the phakir 2024 challenge (2025). 2507.16559.
19. Maier-Hein, L. *et al.* Heidelberg colorectal data set for surgical data science in the sensor operating room. *Sci. Data* **8**, 101 (2021).
20. Roß, T. *et al.* Robust medical instrument segmentation (robust-mis) challenge 2019, 10.7303/SYN18779624 (2019).
21. Budd, C., Garcia-Peraza Herrera, L. C., Huber, M., Ourselin, S. & Vercauteren, T. Rapid and robust endoscopic content area estimation: A lean GPU-based pipeline and curated benchmark dataset. *Comput. Methods Biomech. Biomed. Eng. Imaging & Vis.* **11**, 1215–1224 (2023).
22. Han, Z. *et al.* Robust-mips: A combined segmentation and skeletal representation dataset for surgical instruments in laparoscopic surgery, 10.7303/SYN64023381 (2025).

23. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755 (Springer, 2014).

24. Zheng, C. *et al.* Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.* **56**, 1–37 (2023).

25. Jiang, T. *et al.* RTMPose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399* (2023).

26. Xiao, B., Wu, H. & Wei, Y. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)* (2018).

27. Xu, Y., Zhang, J., Zhang, Q. & Tao, D. VITPose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* **35**, 38571–38584 (2022).

28. MMPose Contributors. OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020).

29. Li, Y. *et al.* SimCC: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, 89–106 (Springer, 2022).

Acknowledgements

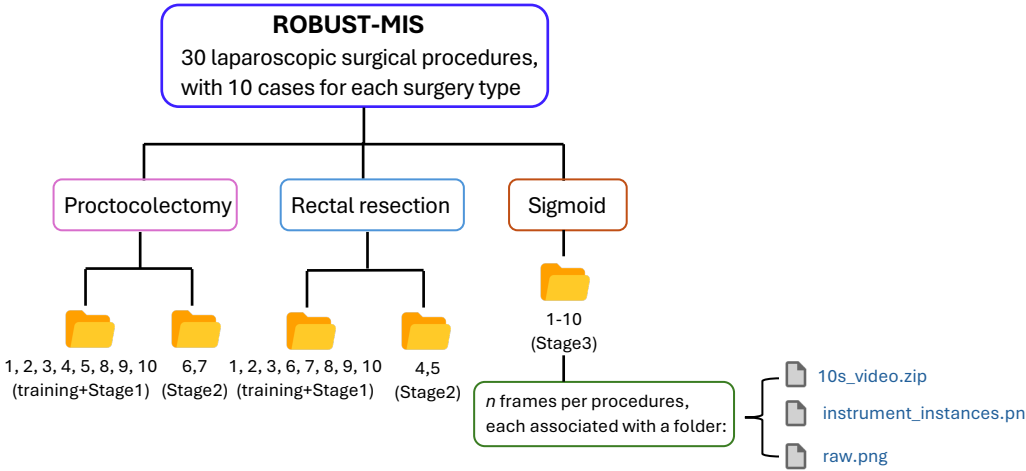
Data Sources: We would like to thank the authors of the ROBUST-MIS dataset for making their data publicly available, which served as the foundation for this work. Funding Sources: This work was supported by core funding from Wellcome/EPSRC [WT203148/Z/16/Z; NS/A000049/1]. Additional support was received from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101016985 (FAROS project), and from Wellcome [WT223880/Z/21/Z]. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

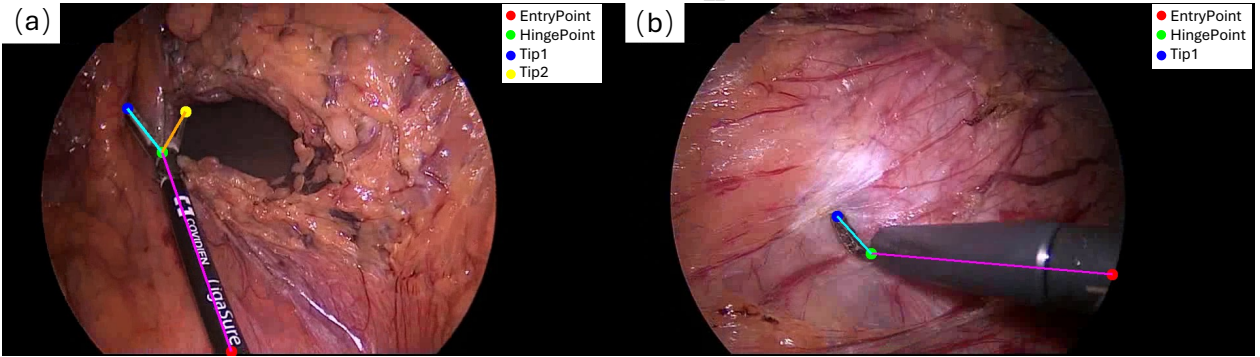
Author Contributions Statement

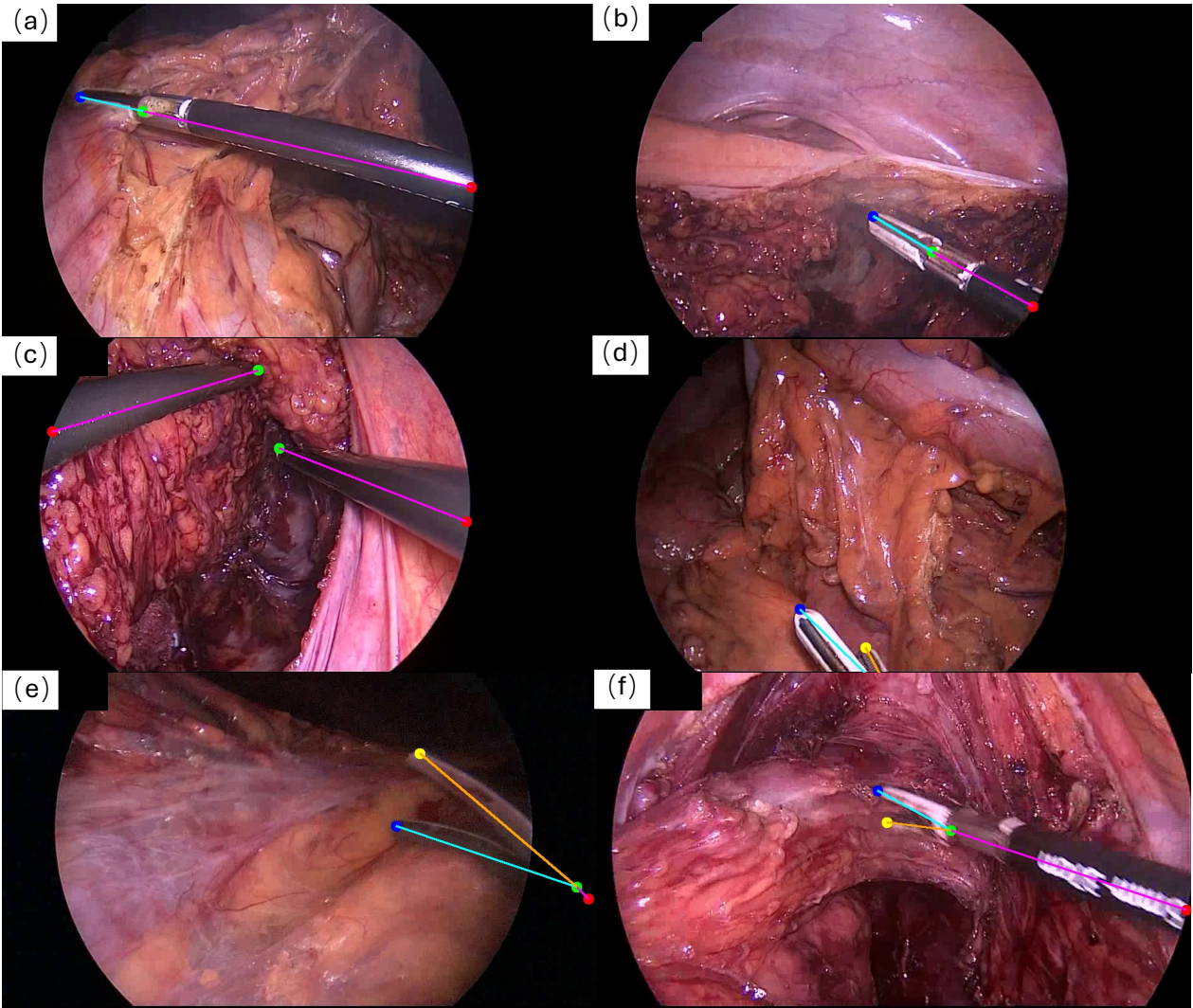
Zhe Han: Data curation, Methodology, Validation, Writing- Original draft preparation. Charlie Budd: Software, Data curation, Writing- Reviewing and Editing. Gongyu Zhang: Writing- Reviewing and Editing. Huanyu Tian: Data curation, Writing- Reviewing and Editing. Christos Bergeles: Supervision. Tom Vercauteren: Conceptualisation, Supervision.

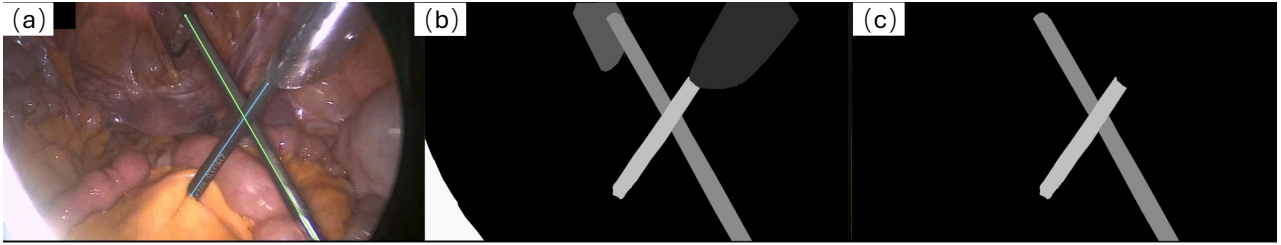
Competing Interests

T.V. is a co-founder and shareholder of Hypervision Surgical Ltd, London, UK. The authors declare that they have no other conflict of interest.






(a)
(b)
(c)
(d)
(e)
(f)

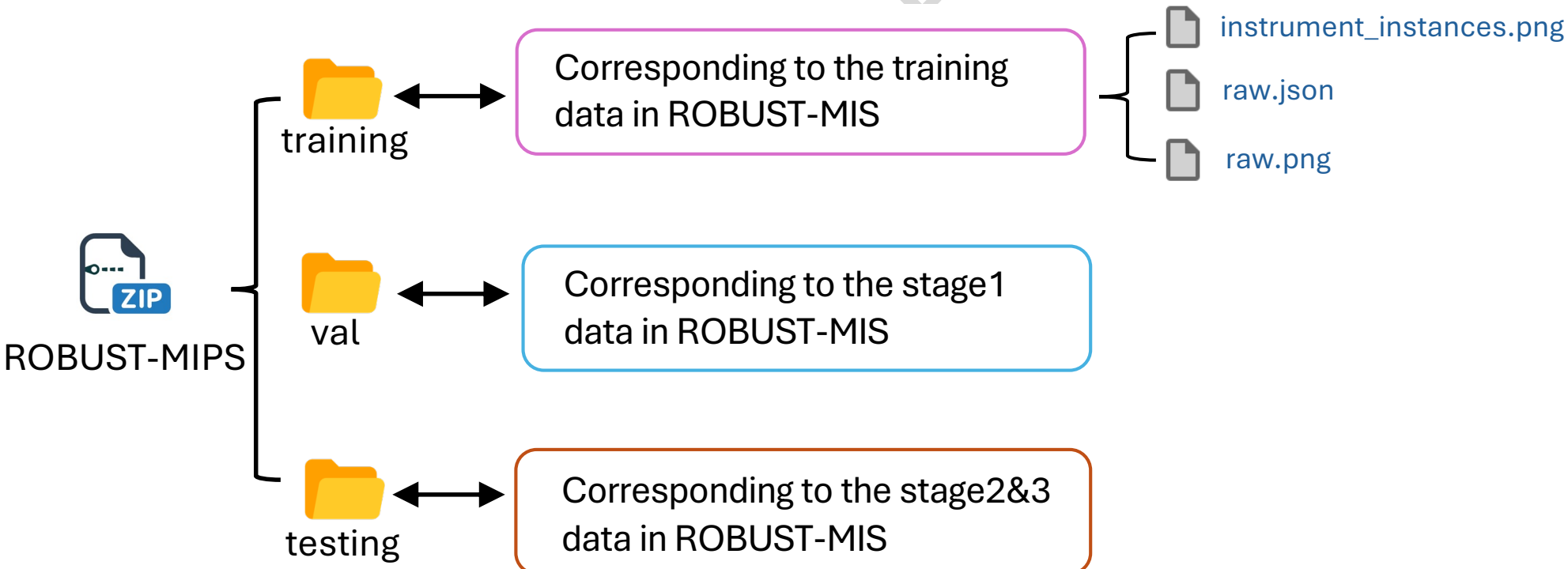


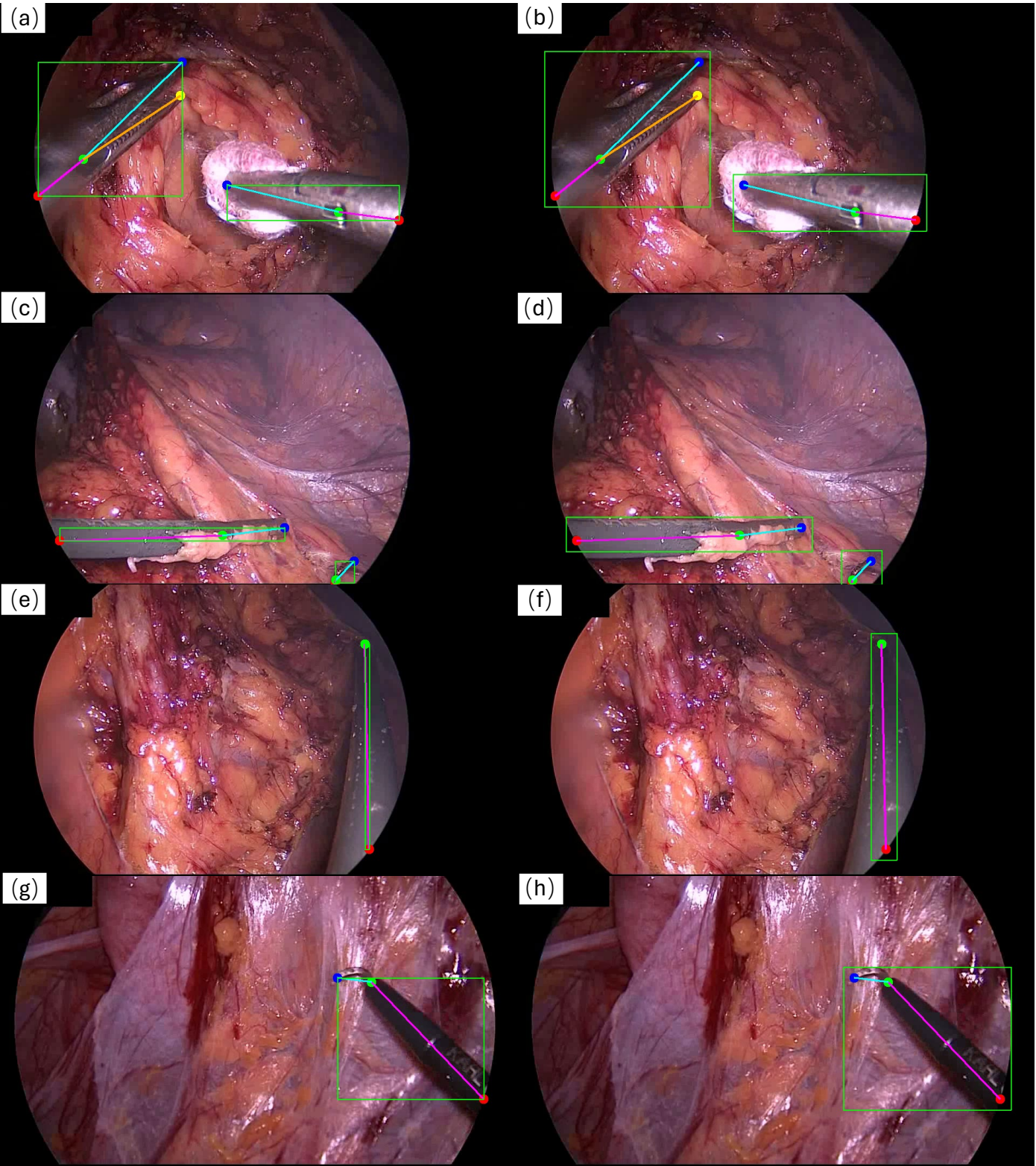
(a)

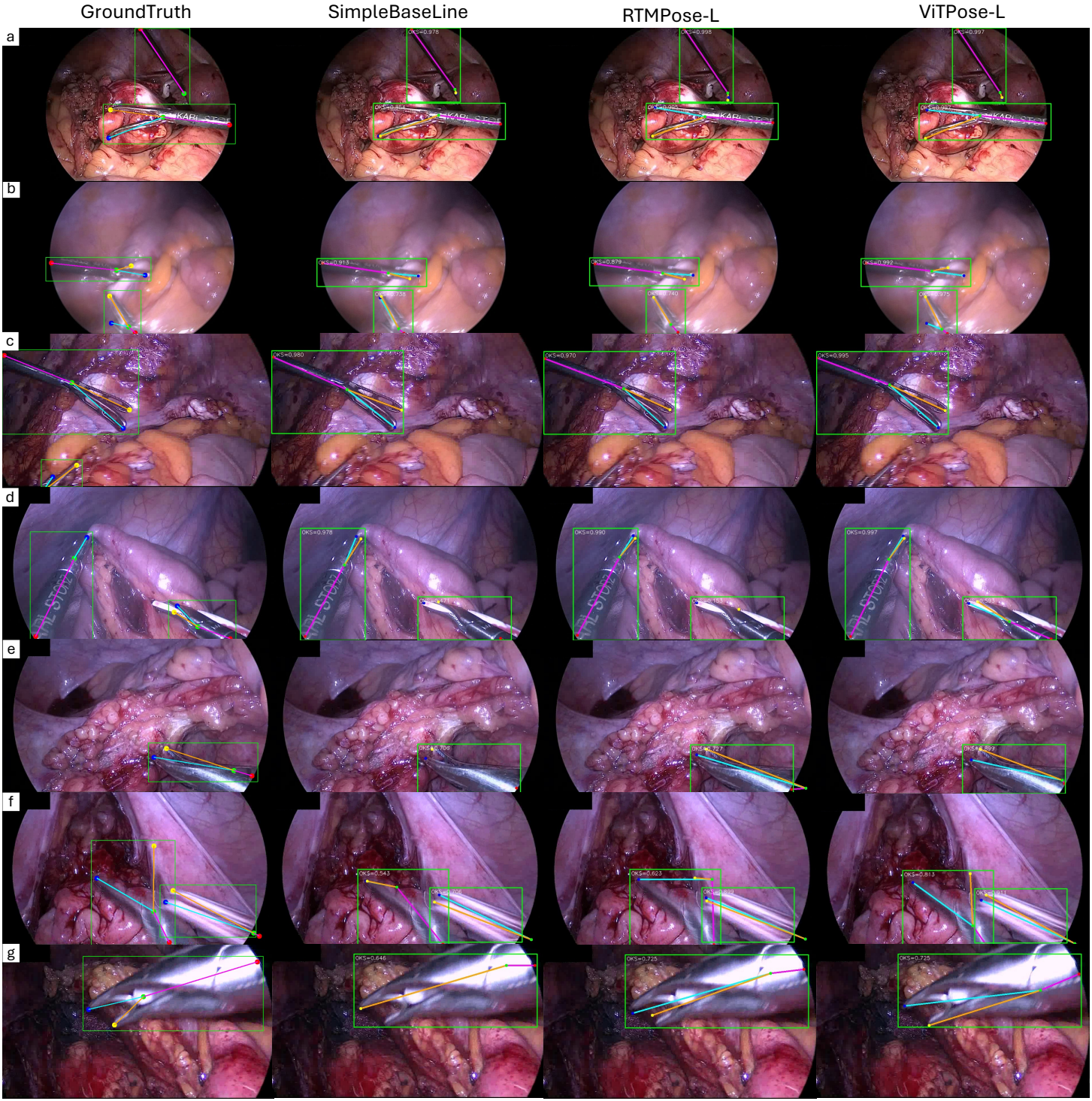

```
[
  {
    "nodes": [
      [247.9, 533.9],
      [208.0, 408.1],
      [149.1, 244.2],
      null
    ],
    "tags": ["visible","visible","visible","missing"],
    "edges": [[0, 1],[1, 2],[1, 3]],
    "transitions": [[],[],[]]
  },
  {
    "nodes": [
      [390.8, 567.9],
      [381.3, 555.0],
      [337.5, 499.6],
      null
    ],
    "tags": ["occluded","occluded","visible","missing"],
    "edges": [[0, 1],[1, 2],[1, 3]],
    "transitions": [[],[[366.6, 536.5]],[]]
  }
]
```


(b)


```
{
  "categories": [
    {
      "supercategory": "SurgicalTool",
      "id": 1,
      "name": "SurgicalTool",
      "keypoints": ["entry", "hinge", "tip1", "tip2"],
      "skeleton": [[0, 1], [1, 2], [1, 3]]
    }
  ],
  "images": [
    {
      "file_name": "file_dir/Stage2_Proctocolectomy_6_1500.png",
      "height": 540,
      "width": 960,
      "id": 0
    }
    ...
  ],
  "annotations": [
    {
      "category_id": 1,
      "image_id": 0,
      "id": 0,
      "bbox": [22, 52, 302, 295],
      "area": 89114.5,
      "keypoints": [42.5, 327.5, 2, 159.2, 219.2, 2,
                    106.7, 72.5, 2, 304.2, 123.3, 2],
      "num_keypoints": 4
    }
    ...
  ]
}
```




(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)


GroundTruth
SimpleBaseLine
RTMPose-L
ViTPose-L
a
b
c
d
e
f
g