VNU – INTERNATIONAL SCHOOL

Pose Estimation of Surgical Instruments using Convolutional Neural Networks for MIS Applications

Author:

Tran Duc Vinh

A thesis submitted in fulfillment of the requirements for the degree of MICE

in the

VNU – International School

June 22, 2025

Declaration of Authorship

I, Tran Duc Vinh , declare that this thesis titled, “Pose Estimation of Surgical Instruments using Convolutional Neural Networks for MIS Applications” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

VNU – INTERNATIONAL SCHOOL

Abstract

Faculty Name
VNU – International School

Pose Estimation of Surgical Instruments using Convolutional Neural Networks for MIS Applications

by Tran Duc Vinh

Accurate detection and pose estimation of surgical instruments are critical for computer-assisted interventions (CAI) and robotic-assisted surgeries. This research proposes an innovative method for detecting and estimating the pose of multiple surgical tools using the YOLOv8-pose model. A comprehensive dataset comprising images of clippers, irrigators, and scissors was meticulously collected and annotated to train the model, facilitating precise localization and orientation estimation of these instruments during laparoscopic procedures.

The performance of the model was assessed using a test dataset across four variants of YOLOv8-pose. Notably, the YOLOv8n variant, characterized by its lightweight architecture with only 3 million parameters, exhibited superior performance in both pose estimation and object detection tasks. For pose estimation, it achieved a Precision of 91%, Recall of 93%, mean Average Precision at IoU 0.5 (mAP@0.5) of 97.9%, and mAP@0.5-0.95 of 88.7%, underscoring its capability to reliably track surgical instruments. In terms of object detection, the model recorded a Precision of 97.9%, Recall of 96.0%, mAP@0.5 of 99.2%, and mAP@0.5-0.95 of 64.6%, demonstrating robust identification and real-time tracking of multiple instruments in surgical settings.

These findings affirm YOLOv8n as an exceptionally efficient model for real-time surgical instrument tracking and pose estimation, rendering it highly suitable for integration into robotic-assisted and minimally invasive surgical systems. Furthermore, this study establishes a foundation for extending the methodology to encompass additional surgical instruments, thereby advancing automation and precision in AI-driven surgical technologies.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AR	Augmented Reality
CIoU	Complete Intersection over Union
CNN	Convolutional Neural Network
DFL	Distribution Focal Loss
GCN	Graph Convolutional Network
IoU	Intersection over Union
mAP	mean Average Precision
MIS	Minimally Invasive Surgery
MSE	Mean Squared Error
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
VR	Virtual Reality
YOLO	You Only Look Once

Chapter 1

Introduction

1.1 Rationale

Recent advancements in computer vision and artificial intelligence (AI) have significantly reshaped numerous industries, with medical technology standing out as a domain of immense potential. Within this realm, pose estimation—defined as the process of ascertaining the position and orientation of an object or tool—has emerged as a pivotal technique, particularly in the field of robotic-assisted surgery. The core objective of pose estimation in surgical applications is to enable precise tracking of instruments during procedures, thereby enhancing procedural safety, accuracy, and efficiency (Hasan et al., 2021).

The escalating complexity of modern surgical interventions has rendered traditional manual tracking methods insufficient for maintaining the precision demanded by contemporary standards. Human factors such as error, fatigue, and the inherent limitations of visual perception in detecting minute movements introduce substantial risks during operations. In this context, pose estimation provides a sophisticated technological solution, ensuring that surgical instruments are consistently maintained in their intended positions and orientations throughout procedures, thus reducing potential errors and bolstering patient safety.

The deployment of pose estimation systems in surgical settings facilitates continuous, real-time monitoring of tools, enabling integration with automated systems to assist surgeons in precise instrument manipulation. Such systems can be seamlessly embedded within robotic surgical platforms, offering enhanced feedback to both robotic mechanisms and human operators. This functionality is especially critical in minimally invasive surgeries (MIS), where precision in tool handling is essential to minimize damage to surrounding healthy tissues (Le et al., 2023).

This study is motivated by the pressing need to advance surgical practices through technological innovation. Specifically, it aims to develop a system capable of detecting and estimating the pose of surgical instruments with high accuracy. Although pose estimation has found widespread application in domains such as manufacturing, augmented reality, and gaming, its integration into the medical field—particularly for surgical tool tracking—remains relatively underexplored. This research seeks to address this gap by investigating the potential of pose estimation to enhance the efficacy and safety of surgical procedures (Hager, Chang, and Morse, 1995).

Furthermore, with the rising prevalence of robotic-assisted surgeries, the provision of accurate, real-time data on surgical tool poses has become increasingly vital. Improved pose detection can mitigate costly procedural errors, ensure proper instrument application, and ultimately elevate patient outcomes. Additionally, this technology offers significant value in medical training, delivering real-time feedback on surgical performance and aiding novice surgeons in mastering correct instrument-handling techniques (Wang et al., 2019).

Ultimately, this research endeavors to establish an efficient and reliable pose estimation system for surgical tools, aiming to revolutionize surgical practices. By contributing to the evolution of cutting-edge medical technologies, this study aims to empower healthcare professionals to deliver safer and more effective treatments, thereby advancing the frontier of AI-driven surgical innovation.

1.2 Aim and Objectives of the Study

1.2.1 Aim

The central aim of this study is to develop and assess a pose estimation system designed for the real-time detection and tracking of surgical instruments, harnessing advanced frameworks in computer vision and machine learning. This research seeks to elevate the precision, efficiency, and safety of minimally invasive surgeries (MIS) and robotic-assisted procedures by enabling accurate determination of tool position and orientation. By advancing pose estimation technologies, the study aspires to contribute to the broader landscape of AI-driven surgical systems, fostering improved procedural outcomes and supporting the evolution of intelligent automation in surgical practice.

1.2.2 Objectives

To achieve this overarching aim, the study outlines a series of objectives that guide the exploration, development, and evaluation of pose estimation frameworks for surgical tools within the context of MIS and robotic-assisted surgeries. These objectives are informed by the theoretical foundations, methodologies, and empirical findings presented throughout the research:

The primary objective is to design and implement a pose estimation system capable of real-time detection and tracking of surgical instruments during operative procedures. Drawing from the technological insights in Chapter 2, this system will leverage camera-based solutions and machine learning frameworks to ensure precise localization and orientation estimation, enhancing surgical precision and patient safety across dynamic operative environments.

A key goal is to compile and preprocess a robust dataset to support the training and validation of pose estimation models, as elaborated in Chapter 3. This dataset will encompass annotated endoscopic images of various surgical tools—such as clippers, irrigators, and scissors—captured under diverse conditions, with augmentation techniques applied to improve model adaptability and generalization to real-world surgical scenarios.

Another essential objective is to evaluate the performance of pose estimation frameworks, focusing on their accuracy, processing speed, and robustness, as evidenced by the experimental results in Chapter 4. The study will assess these systems against standardized metrics, including precision, recall, and mean average precision (mAP), to identify optimal approaches for real-time tracking of surgical instruments in MIS contexts.

The research also seeks to compare the developed pose estimation system with established techniques in the field, as discussed in Chapter 4. This comparative analysis will explore a range of frameworks—spanning traditional methods like marker-based tracking to modern deep learning approaches—to highlight their respective strengths, limitations, and suitability for surgical applications, emphasizing computational efficiency and practical deployment.

A further objective is to investigate the practical utility of pose estimation systems in real-world surgical settings, as outlined in Chapter 4. This includes their integration with robotic surgery platforms to provide real-time positional feedback, thereby enhancing surgeon control and procedural accuracy. Additionally, the study aims to explore their role in surgical training by enabling movement tracking and performance assessment, fostering skill development among practitioners.

Finally, the study aims to delineate future research directions to expand the capabilities of pose estimation technologies, as proposed in Chapters 4 and 5. This encompasses broadening the scope to additional surgical instruments, incorporating 3D pose estimation through depth information, integrating augmented reality (AR) for enhanced visualization, and exploring advanced learning paradigms—such as semi-supervised methods—to reduce dependency on extensive manual annotations. These efforts seek to lay a foundation for increased automation and autonomy in future surgical systems.

Through these objectives, this research endeavors to deliver a versatile, efficient, and robust pose estimation framework that enhances the efficacy of robotic-assisted and minimally invasive surgical procedures, while establishing a platform for ongoing advancements in AI-supported medical technologies.

1.3 Research questions

The research aims to address several key questions that will guide the development, implementation, and evaluation of the pose estimation system for surgical tools. These questions focus on both the technical aspects of the system and its practical application in a surgical environment. The primary research questions are as follows:

1. How accurate and reliable can the pose estimation system be in detecting and tracking the position and orientation of surgical tools in real-time?

2. Which pose estimation algorithms provide the best balance between accuracy, speed, and robustness when applied to the tracking of surgical tools?

3. What type of data (e.g., images, sensor data, etc.) and camera setup are most effective for training and operating the pose estimation system in a real-world surgical setting?

4. How can the pose estimation system be integrated into a robotic surgery platform to improve surgical precision and provide real-time feedback to surgeons?

5. What are the challenges and limitations of using pose estimation for surgical tool tracking, and how can they be addressed?

6. Can the pose estimation system be scaled or adapted for use with a variety of surgical tools and in different types of surgeries?

7. What are the potential benefits and drawbacks of using pose estimation in real-time surgical tool tracking, from both a clinical and a technological perspective?

1.4 Methods of the Study

This study employs a blend of qualitative and quantitative research methodologies to investigate pose estimation frameworks for surgical tool tracking, aiming to address the outlined research questions and objectives. The approach integrates multiple stages, from theoretical groundwork to practical evaluation, ensuring a comprehensive exploration of the system's development and application in surgical contexts.

The research begins with a systematic literature review to synthesize existing knowledge on pose estimation, computer vision, and surgical instrument tracking. This step establishes a theoretical foundation by identifying current technologies, methodologies, and gaps in the field, particularly within minimally invasive surgery (MIS) and robotic-assisted procedures, which informs subsequent system design and development efforts.

The development of the pose estimation system involves designing a framework tailored for real-time detection and tracking of surgical tools. This process includes selecting suitable algorithms—spanning traditional and deep learning-based approaches—and constructing a software platform capable of processing endoscopic data. A diverse dataset of surgical tool images featuring instruments like clippers, irrigators, and scissors is meticulously curated from real-world surgical videos. This dataset is preprocessed and annotated with keypoints to provide ground truth, enhancing the system's ability to generalize across varied surgical scenarios.

Training and testing form a core component of the methodology, where the collected dataset is utilized to refine machine learning models optimized for pose estimation tasks. Convolutional neural networks (CNNs) and other advanced architectures are trained, with hyperparameters tuned to balance accuracy and computational efficiency. The trained system is then rigorously tested on unseen data to evaluate its real-time tracking performance and adaptability to dynamic conditions such as occlusions and variable lighting.

Evaluation of the system's effectiveness is conducted through a suite of performance metrics, including accuracy, speed, and robustness, under simulated surgical conditions. A comparative analysis benchmarks the proposed framework against existing pose estimation techniques, highlighting its strengths in precision and efficiency. The system is further integrated with a robotic surgery platform to assess its capacity to deliver real-time feedback, enhancing surgical precision and decision-making in practical settings.

Qualitative insights complement the quantitative findings by gathering feedback from medical professionals and surgeons on the system's usability and potential impact on surgical practices. This dual approach ensures a holistic understanding of both technical performance and clinical applicability, addressing challenges like data dependency and computational demands while identifying opportunities for real-world deployment and future enhancements.

1.5 Scope of the Study

This study centers on the development and evaluation of a pose estimation system tailored for surgical instruments, with the overarching goal of enhancing the precision and safety of robotic-assisted surgery. The research delves into the application of advanced computer vision and machine learning techniques, particularly deep learning models such as convolutional neural networks (CNNs), to enable real-time detection and tracking of surgical tool positions and orientations. By focusing on these technological domains, the study seeks to establish a robust framework capable of addressing the dynamic demands of surgical environments, thereby contributing to the broader advancement of automated surgical technologies.

The investigation targets a selection of commonly utilized surgical tools, including clippers, irrigators, and scissors, which are integral to minimally invasive procedures. The primary emphasis lies in monitoring the spatial positioning and orientation of these instruments during surgical operations, ensuring accurate tracking within the operative field. This focus is grounded in the need to provide reliable data that can support surgical precision, particularly in scenarios where manual oversight alone may prove insufficient.

The pose estimation system is developed and assessed within a simulated surgical environment, utilizing recorded endoscopic videos and imagery as the primary data source. This controlled setting serves as the foundation for training and validating the system, allowing for an initial evaluation of its performance before potential application in more realistic or clinical contexts. The reliance on simulation reflects a pragmatic approach to testing under replicable conditions, facilitating a comprehensive analysis of the system's capabilities across a spectrum of surgical scenarios.

Evaluation of the system encompasses a thorough assessment of its accuracy, processing speed, and robustness, conducted under diverse conditions such as variable lighting, occlusions, and differing tool configurations. These performance metrics are critical to determining the system's efficacy in real-time tracking, a cornerstone of its intended utility in robotic-assisted surgery. Additionally, the research explores the feasibility of integrating the pose estimation framework with existing robotic surgery platforms, with a particular interest in its potential to deliver actionable real-time feedback to surgeons, thereby enhancing procedural decision-making and control.

The scope of this study is deliberately bounded, excluding the design or development of robotic surgery systems themselves, as well as the direct implementation of the pose estimation framework in live clinical settings. Instead, the focus remains on the conceptualization, implementation, and evaluation of the pose estimation system, with future investigations proposed to bridge the gap toward practical surgical applications. The research is further constrained by the availability and diversity of training data, which may limit the system's ability to generalize across all possible surgical conditions and tool types. Moreover, the study prioritizes tools and procedures typical of standard medical practice, eschewing exploration into highly specialized or complex surgical interventions.

1.6 Main Contributions

This thesis advances the field of surgical instrument pose estimation through deep learning by presenting the following key contributions:

(1) **YOLOv8-Pose Based System for Surgical Tool Detection:** This study develops and evaluates a real-time system using YOLOv8-Pose for detecting and estimating the pose of surgical tools—clippers, irrigators, and scissors—in laparoscopic surgery. It analyzes the model's architecture, training, and performance, highlighting its precision and efficiency for intraoperative use.

(2) **Curated Dataset and Annotation:** A specialized dataset of laparoscopic surgery images, derived from the m2cai16-tool-locations dataset and enriched with additional sources, was curated and annotated with precise keypoints for multiple surgical tools. Utilizing the Computer Vision Annotation Tool (CVAT), this meticulously documented resource enhances training and benchmarking of pose estimation models, addressing a critical need in medical AI research.

(3) **Comparative Analysis:** A thorough comparison of YOLOv8-Pose with YOLOv5-Pose, HRNet, and OpenPose demonstrates its superiority in accuracy, speed, and computational efficiency for surgical tool tracking, especially with the lightweight YOLOv8n variant in real-time medical applications.

(4) **Real-Time Performance and Practical Applicability:** The study showcases YOLOv8-Pose's real-time capabilities, enabling seamless integration into robotic-assisted surgery, augmented reality navigation, and surgical training systems. Its practical implications for enhancing minimally invasive procedures are substantiated through optimized performance on embedded devices like Jetson Nano.

(5) **Published Research:** A core component focusing on clipper tool detection and pose estimation was peer-reviewed and published in the 2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV), cited as Kim et al., "Surgical Tool Detection and Pose Estimation using YOLOv8-pose Model: A Study on Clipper Tool," pp. 225-229, IEEE, 2024, affirming the work's novelty and impact within the research community.

Chapter 2

Minimally Invasive Surgery and Pose Estimation

2.1 Minimally Invasive Surgery

Minimally Invasive Surgery (MIS) has emerged as a transformative approach in modern medicine, redefining surgical practices by prioritizing reduced invasiveness over traditional open techniques. Unlike conventional surgery, which relies on large incisions to access internal organs, MIS utilizes small incisions—commonly referred to as "keyhole surgery"—through which specialized instruments and endoscopic cameras are deployed. This method substantially diminishes tissue trauma, alleviates postoperative pain, reduces infection risks, and expedites recovery, ultimately lowering hospital stays and healthcare costs (Fuchs, 2005; Mack, 2001). The historical milestone of laparoscopic cholecystectomy in the late 1980s marked the onset of MIS's widespread adoption, extending its application across specialties such as gynecology, urology, orthopedics, and general surgery, thereby enhancing patient outcomes through minimized morbidity and improved cosmetic results (**Reynolds2001**; Dubois et al., 1990).

Despite its advantages, MIS poses significant challenges for surgeons. The reliance on two-dimensional endoscopic imagery restricts the field of view and impairs depth perception, while the absence of direct haptic feedback—integral to traditional surgery—complicates precise manipulations. Surgeons must adapt through enhanced hand-eye coordination and specialized training to navigate these limitations effectively (Okamura, 2009; Gallagher et al., 2003). These inherent difficulties underscore the need for technological innovations to bolster MIS's efficacy, ensuring its benefits are fully realized in clinical practice.

To address these challenges, advanced technologies have been integrated into MIS workflows, significantly enhancing surgical capabilities. Computer vision plays a pivotal role by enabling real-time recognition and tracking of surgical instruments, thereby improving visualization and procedural monitoring. This technology employs sophisticated algorithms to ascertain tool positions and orientations, offering critical feedback to surgeons (Maier et al., 2019). Similarly, artificial intelligence (AI) augments surgical decision-making through predictive analytics and automation of routine tasks, such as suturing, while analyzing video streams to detect anomalies and optimize strategies (Hashimoto et al., 2018). Robotic-assisted systems, epitomized by platforms like the da Vinci Surgical System, further elevate precision and dexterity, allowing surgeons to execute complex procedures with reduced error and tremor (Lanfranco et al., 2004).

Complementing these advancements, augmented reality (AR) overlays real-time data—such as 3D anatomical models and critical structure highlights—onto the surgeon's visual field, enhancing navigational accuracy and procedural safety (Meola et al., 2017). Virtual reality (VR) and simulation training provide immersive, risk-free environments for skill development, proving indispensable for mastering intricate MIS techniques and assessing surgical proficiency (McGaghie et al., 2010). Additionally, haptic feedback devices aim to restore tactile sensation through advanced sensors and actuators, enabling surgeons to perceive resistance and texture, thus rendering tool and tissue manipulation more intuitive and potentially minimizing unintended trauma (Kucuk et al., 2016). Together, these technologies synergistically enhance the precision, safety, and training associated with MIS, paving the way for its continued evolution in surgical practice.

2.2 Surgical Instruments and Pose Estimation

Laparoscopic surgery necessitates the use of specialized instruments engineered for precision and adaptability, designed to operate through minimal incisions within confined anatomical spaces. These tools must deliver high maneuverability and control, as surgeons rely predominantly on visual feedback from endoscopic cameras rather than direct observation. The capacity to manipulate these instruments with accuracy is paramount to the success of surgical interventions, ensuring minimal trauma to surrounding tissues while achieving the intended procedural outcomes (Lim and Erdman, 2003).

The array of instruments employed in laparoscopic surgery is tailored to perform intricate tasks within the abdominal cavity. Commonly utilized tools include graspers, which facilitate the holding, manipulation, and stabilization of tissues or organs, and scissors, crafted for cutting tissues, sutures, or fibrous structures, often equipped with curved or angled tips to enhance maneuverability. Clippers are integral for clamping or ligating blood vessels and ducts, deploying small metal clips to prevent bleeding, a function critical in procedures such as cholecystectomies. Hooks, whether sharp or blunt-tipped, serve to lift, separate, or reposition tissues, thereby improving visibility and access to surgical sites. Additionally, irrigators play a vital role by flushing the operative field with sterile fluids, clearing debris and maintaining visibility, which reduces infection risks and supports procedural precision (Lim and Erdman, 2003).

Pose estimation, defined as the determination of a surgical tool's precise position and orientation, is indispensable in laparoscopic and robotic-assisted surgeries, where accurate instrument tracking underpins successful outcomes. This technology enhances surgical precision by providing exact spatial data, enabling surgeons to execute movements with minimal risk of damaging adjacent healthy tissues—a necessity in the constrained visibility of minimally invasive procedures. It also bolsters surgeon control, offering stability for complex tasks such as suturing or dissection, where fine motor skills are essential, and mitigates hand tremors through real-time positional insights. Furthermore, pose estimation delivers continuous feedback, improving situational awareness and decision-making by displaying tool positions on monitors, a feature particularly valuable when integrated with augmented reality (AR) or AI-driven guidance systems to optimize visualization and reduce cognitive demands. Beyond manual enhancement, it lays the groundwork for automation in robotic surgery, enabling AI systems to perform tasks like precise suturing or retraction, thus alleviating surgeon fatigue and enhancing procedural reproducibility (Maier et al., 2019; Weber et al., 2018).

To be effective in the dynamic environment of laparoscopic surgery, a pose estimation system must fulfill stringent requirements. It needs to accurately compute the three-dimensional position and orientation of each instrument, achieving sub-millimeter precision to safeguard delicate anatomical structures during intricate maneuvers. Real-time processing is equally critical, as any latency could jeopardize safety and efficacy by delaying feedback essential for immediate adjustments. The system must also exhibit robustness against challenges such as occlusions from tissues or other tools, variable lighting conditions, and reflections from metallic surfaces, ensuring consistent performance despite these adversities. Compatibility with a diverse range of instruments—varying in shape, size, and function—is necessary to avoid frequent recalibration, while a minimal computational footprint ensures practicality for deployment on embedded surgical platforms. Additionally, the system must adeptly track rapid and complex surgical movements, resisting disruptions from motion blur or abrupt shifts to support the precise articulation demanded in robotic-assisted procedures (Maier et al., 2019; Xu and Giannarou, 2024).

The application of pose estimation in laparoscopic surgery encounters several inherent challenges due to the constrained and dynamic nature of the abdominal cavity. The limited field of view provided by endoscopic cameras restricts tracking when instruments exit the visible range, potentially disrupting positional continuity and complicating trajectory estimation. Occlusions, caused by overlapping tissues, tools, or camera obstructions, further hinder consistent tracking, necessitating advanced strategies like temporal analysis or depth modeling to maintain accuracy. Moreover, variable lighting within the body cavity—exacerbated by glare, shadows, and uneven illumination—introduces noise that can impair image quality and feature extraction, posing additional hurdles to reliable pose computation. These challenges underscore the need for sophisticated preprocessing and adaptive algorithms to ensure robust performance in real-world surgical settings (Xu and Giannarou, 2024).

2.3 Pose Estimation Methods

Pose estimation holds a pivotal role across various healthcare applications, ranging from surgical assistance and medical imaging to patient monitoring, where precise spatial awareness is paramount (Maier et al., 2019). In the domains of laparoscopic surgery and surgical robotics, the ability to determine the position and orientation of instruments in real time is vital for enhancing procedural accuracy and safeguarding patient outcomes (Pedram et al., 2016). The methodologies employed in pose estimation can be broadly classified into traditional techniques, which often depend on external aids, and modern deep learning-based approaches that leverage advanced computational models to interpret visual data directly.

2.3.1 Traditional Methods in Medical Pose Estimation

Traditional pose estimation methods in medical contexts typically rely on external markers or sensors affixed to the objects being tracked, offering reliable solutions in controlled settings but facing limitations in the unpredictable dynamics of surgical environments. Marker-based tracking involves attaching physical markers—often reflective or distinctly colored—to surgical instruments or patient anatomy, which are then monitored by cameras to provide precise three-dimensional localization

(Speidel et al., 2006). While this approach excels in accuracy and ease of integration under stable lighting, its effectiveness diminishes when occlusions from tissues, blood, or other tools disrupt visibility. Optical tracking, another prevalent technique, employs infrared cameras to detect reflective markers, delivering high-precision data crucial for systems like the da Vinci robotic platform (Groeger, Arbter, and Hirzinger, 2008). However, its sensitivity to environmental factors such as intense lighting or obstructions, coupled with the necessity for an unobstructed line of sight, can constrain its utility in complex surgical scenarios. Electromagnetic (EM) tracking utilizes small sensors within a magnetic field to ascertain instrument positions, circumventing the need for direct visibility and proving advantageous in obscured surgical regions (Lugade et al., 2015). Yet, its susceptibility to interference from metallic objects—common in operating rooms—can compromise accuracy. Mechanical tracking, by contrast, integrates sensors into robotic arms or linkages to directly measure tool movements, ensuring exceptional precision and immunity to external conditions like lighting (Lugade et al., 2015). Nevertheless, its reliance on physical constraints limits motion range and flexibility, posing challenges in hybrid or manual procedures, alongside ongoing maintenance demands.

2.3.2 Modern Deep Learning-Based Methods

The advent of deep learning has transformed pose estimation by introducing markerless techniques capable of discerning complex patterns from raw image data, offering enhanced robustness against occlusions and lighting variations compared to traditional methods. Convolutional Neural Networks (CNNs) stand out as a cornerstone of this revolution, employing multiple convolutional layers to extract spatial features and predict keypoints with high accuracy from medical imagery (O'Shea and Nash, 2015). Widely applied in tracking surgical instruments and analyzing patient movements in rehabilitation, CNNs demand substantial labeled datasets and computational resources, which can hinder real-time deployment on resource-constrained systems. Recurrent Neural Networks (RNNs), designed for sequential data processing, excel in tracking motion trajectories and anticipating future movements by retaining contextual memory from prior inputs (Xu et al., 2022). In healthcare, they support applications like patient monitoring and surgical action prediction, though challenges such as the vanishing gradient problem and lengthy training periods persist, mitigated in part by advanced variants like LSTMs. Transformer models, leveraging self-attention mechanisms, capture long-range dependencies across entire images or video frames, surpassing CNNs in processing global spatial relationships (Doughty and Ghugre, 2022). Their use in surgical video analysis and diagnostic imaging underscores their potential, albeit at the cost of significant computational overhead and data requirements. Graph Convolutional Networks (GCNs), meanwhile, model keypoints as interconnected graph structures, adeptly learning spatial dependencies for precise motion tracking in contexts like physical therapy and surgical assistance (Wang et al., 2019). Despite their accuracy, GCNs' computational complexity and sensitivity to noisy data present notable challenges.

2.3.3 Key Considerations for Pose Estimation in Healthcare

Selecting an optimal pose estimation method for healthcare applications necessitates a careful evaluation of multiple factors to ensure its efficacy and practicality. Accuracy remains paramount, as even minor deviations in tracking can lead to critical errors in surgical tool alignment or diagnostic interpretation, directly impacting patient safety. Processing time is equally critical, with real-time performance being essential to synchronize tracking with surgical actions, requiring low-latency computation to maintain operational flow (Pedram et al., 2016). Robustness to occlusions is a vital consideration, as partial obscurations by tissues or instruments are commonplace in surgery, necessitating techniques like temporal tracking or depth inference to sustain reliability (Shotton et al., 2011). Scalability is another key aspect, demanding that the method seamlessly integrate with diverse surgical systems and adapt to a variety of instruments, enhancing its versatility across medical contexts. Finally, computational resource demands must align with available hardware, favoring lightweight models that enable efficient deployment on embedded devices or robotic platforms without sacrificing performance (Krizhevsky, Sutskever, and Hinton, 2017). These considerations collectively guide the choice of pose estimation techniques, balancing precision with practical applicability in healthcare settings.

2.4 Deep Learning Models for Pose Estimation

Deep learning has fundamentally transformed pose estimation by enabling precise and efficient determination of object position and orientation within intricate environments, marking a significant leap forward in computational capabilities. In healthcare, particularly within surgical contexts and medical imaging, these techniques have revolutionized the tracking of surgical instrument movements, the comprehension of anatomical structures, and the enhancement of patient monitoring systems (Maier et al., 2019). By leveraging advanced neural network architectures, deep learning models extract complex patterns directly from raw data, offering robust solutions that adapt to the dynamic demands of medical applications. This section delves into key deep learning models employed in pose estimation—Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs), and hybrid approaches—highlighting their relevance and utility in healthcare settings.

2.4.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are engineered to process grid-like data, such as medical images, through a series of convolutional layers that apply filters to extract spatial hierarchies of features, progressing from rudimentary elements like edges and textures to sophisticated representations of anatomical or instrumental forms (LeCun et al., 1998). This automated feature extraction, bolstered by pooling layers that reduce dimensionality and fully connected layers that interpret these features, empowers CNNs to discern patterns with remarkable efficacy, eliminating the need for manual feature engineering (Krizhevsky, Sutskever, and Hinton, 2017). In healthcare, CNNs are instrumental in real-time tracking of surgical tools during laparoscopic and robotic-assisted procedures, ensuring precise control and minimizing errors by recognizing instruments amidst complex visual scenes (Allan et al., 2020). Beyond surgery, they excel in medical image analysis, detecting tumors and lesions in MRI, CT, and X-ray scans to facilitate early diagnosis of conditions like cancer, while also monitoring patient posture in rehabilitation to assess therapeutic progress (Litjens et al., 2017). Their strengths lie in high precision, resilience to noise and occlusions, and the ability to leverage transfer learning from pre-trained models, mitigating the demand for extensive labeled medical datasets (Pan and Yang, 2010). However, CNNs require significant computational resources, often necessitating GPUs or TPUs, and can suffer from overfitting on limited data, compounded by their opaque "black box" nature, which challenges interpretability.

2.4.2 Graph Convolutional Networks (GCNs)

Graph Convolutional Networks (GCNs) offer a specialized approach to pose estimation by modeling data as graph structures, where nodes represent keypoints—such as joints, landmarks, or tool tips—and edges delineate their spatial relationships (Kipf and Welling, 2016). Unlike CNNs, which are constrained to Euclidean grid data, GCNs extend convolution to non-Euclidean spaces, aggregating information from neighboring nodes to capture intricate structural patterns and dependencies (Wu et al., 2020). In medical applications, GCNs prove invaluable for analyzing patient joint movements during rehabilitation, enabling personalized treatment adjustments by tracking progress with high fidelity (Yan, Xiong, and Lin, 2018). They also excel in surgical tool pose estimation, modeling the spatial interplay of instrument components for real-time orientation tracking in minimally invasive procedures (Wang et al., 2019). Additionally, GCNs map neural connectivity in fMRI to study neurological disorders or analyze vascular networks in cardiovascular imaging, offering insights into disease mechanisms (Parisot et al., 2017). Their advantages include adept handling of irregular data and enhanced interpretability through graph representations, yet they demand significant computational power and are sensitive to noisy inputs, with effective graph construction often requiring specialized domain knowledge.

2.4.3 Hybrid Deep Learning Models

Hybrid deep learning models integrate multiple neural network architectures to harness their complementary strengths, delivering superior performance in the multifaceted challenges of healthcare-related pose estimation (Zhang et al., 2020). These models typically combine CNNs, which excel at spatial feature extraction from visual data, with architectures like Recurrent Neural Networks (RNNs) or Transformers to process sequential data and capture temporal dynamics, often incorporating GCNs to model spatial relationships in non-Euclidean contexts. In surgical settings, such hybrids enable real-time tool tracking by merging CNN-derived visual features with RNN-analyzed movement sequences, ensuring dynamic precision in robotic-assisted procedures. They also support comprehensive patient monitoring in intensive care units, where CNNs process imaging data alongside Transformers interpreting time-series vital signs, facilitating early detection of deterioration. For multi-modal medical image analysis, integrating GCNs, CNNs, and Transformers enhances diagnostic accuracy across diverse imaging types, while in rehabilitation, these models track patient progress over time, adapting treatment plans accordingly. Furthermore, hybrid approaches model disease progression by combining GCN-mapped biological networks with Transformer-analyzed patient histories, paving the way for personalized medicine through predictive insights (Wang et al., 2019). These models offer robust predictive power, resilience to noise, and adaptability to heterogeneous data, supporting real-time analysis in dynamic healthcare environments, though their complexity increases computational demands and integration challenges.

2.5 YOLO Model and Its Applications

The You Only Look Once (YOLO) model represents a groundbreaking advancement in real-time object detection within computer vision, profoundly influencing a wide array of applications, including those in healthcare (Redmon et al., 2016). Renowned for its exceptional speed and accuracy, YOLO's ability to simultaneously detect and localize multiple objects within a single image has positioned it as an ideal solution for medical scenarios where swift and precise identification is paramount (Zhao et al., 2019). By processing visual data in a streamlined manner, YOLO offers a robust framework that meets the rigorous demands of real-time medical diagnostics and surgical interventions, establishing its significance across diverse healthcare domains.

2.5.1 Mechanism of YOLO

Distinct from conventional object detection approaches that rely on region proposal networks or sliding windows, YOLO employs a single forward pass through a convolutional neural network to analyze an entire image, enhancing its computational efficiency (Redmon and Farhadi, 2017). The input image is segmented into an S x S grid, with each cell tasked with predicting multiple bounding boxes, each accompanied by a confidence score that reflects the probability of an object's presence within that region. Simultaneously, each grid cell estimates conditional class probabilities, indicating the likelihood that an object, if present, belongs to a specific category—such as surgical tools like clippers or irrigators (Redmon and Farhadi, 2018). This unified architecture culminates in rapid detection by thresholding these confidence scores and applying non-maximum suppression to eliminate redundant boxes, rendering YOLO exceptionally suited for applications requiring immediate processing, such as surgical navigation and diagnostics.

2.5.2 Applications in Healthcare

YOLO's remarkable speed and precision have catalyzed its widespread adoption across various healthcare applications, enhancing both clinical efficiency and patient outcomes. In the realm of minimally invasive surgeries (MIS), including laparoscopic and robotic-assisted procedures, YOLO facilitates real-time detection and tracking of surgical instruments, bolstering precision and minimizing human error to improve safety (Schoenthaler, Lassner, and Gühmann, 2020). Beyond surgery, its prowess extends to medical imaging, where it swiftly identifies abnormalities—such as tumors, cysts, or fractures—across modalities like X-rays, MRI, and CT scans, enabling early diagnosis and intervention for critical conditions (Agarwal, Goel, and Gupta, 2020). In endoscopy, YOLO processes video feeds to detect gastrointestinal anomalies like polyps or ulcers in real time, augmenting diagnostic accuracy and reducing oversight (Misawa et al., 2018). Additionally, it supports patient monitoring in intensive care units by recognizing distress signals or unusual movements, and aids fall detection for vulnerable populations, enhancing care delivery (Ravanelli et al., 2018). In radiology and pathology, YOLO streamlines workflows by automating the detection of radiological and histopathological abnormalities, prioritizing urgent cases and accelerating diagnostic processes (Cireșan, Meier, and Schmidhuber, 2012).

2.5.3 Advantages of YOLO

YOLO's design confers several advantages that align seamlessly with the demands of real-time medical applications. Its single-pass architecture ensures rapid processing, making it an invaluable asset for time-sensitive tasks like surgical assistance and emergency diagnostics. Despite its emphasis on speed, YOLO maintains competitive accuracy in detecting multiple objects concurrently, ensuring reliable performance across complex scenes. The end-to-end detection process simplifies computational workflows, enhancing efficiency compared to multi-stage methods (Bochkovskiy, Wang, and Liao, 2020). Its versatility allows adaptation to diverse medical imaging modalities and tasks, while its efficient resource utilization—requiring fewer computational demands than some advanced models—enables deployment on edge devices, broadening its practical reach. Furthermore, YOLO's scalability and customization potential, achievable through fine-tuning on specialized datasets, make it a flexible tool for tailored healthcare solutions.

2.5.4 Limitations of YOLO

Despite its strengths, YOLO is not without limitations that warrant consideration in healthcare contexts. It often struggles to detect small objects that occupy minimal portions of a grid cell, a challenge particularly relevant in medical imaging where subtle lesions or fine instruments may be overlooked (Redmon and Farhadi, 2018). Localization errors can also arise, especially in scenarios with overlapping or densely packed objects, potentially compromising precision in crowded surgical fields. The model's reliance on large, annotated datasets for effective training poses a hurdle in healthcare, where data privacy and annotation costs limit availability, necessitating extensive resources for optimal performance. Generalization issues further complicate its use, as performance may degrade on datasets diverging significantly from training conditions, requiring domain-specific adjustments. Additionally, like many deep learning models, YOLO's opaque decision-making process hinders interpretability, presenting challenges for clinical validation and trust in critical applications (Zhao et al., 2019).

Chapter 3

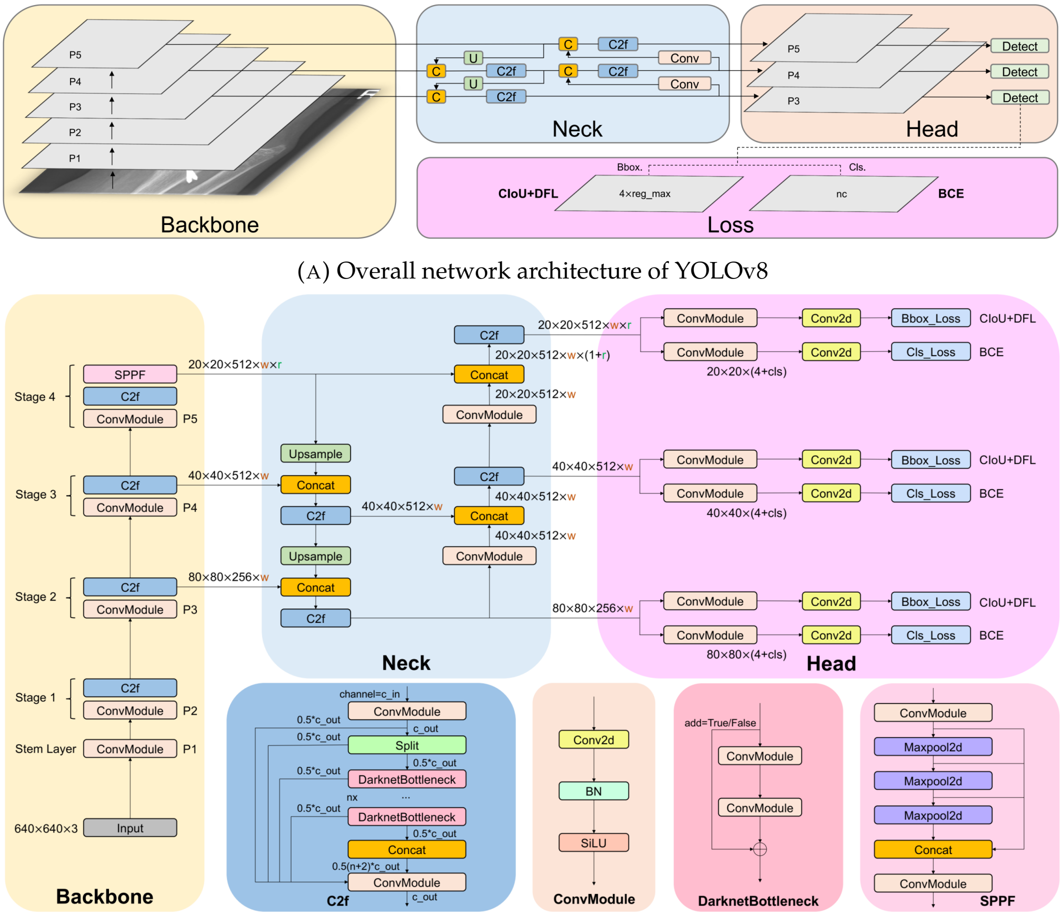
Methodology

3.1 YOLOv8 Model

Object detection stands as a pivotal domain within computer vision, with its utility spanning numerous applications across various fields. Within this landscape, the YOLO (You Only Look Once) family of models has emerged as a trailblazer in real-time object detection, continuously advancing the benchmarks for speed and accuracy. This section details YOLOv8, the latest iteration released by Ultralytics, highlighting its architectural innovations, loss function formulation, and its extension to pose estimation, with a particular focus on its application to surgical tool tracking in laparoscopic surgery. While a formal peer-reviewed paper specifically detailing YOLOv8's internal workings is not yet available (as of the original document's date), we leverage available documentation, code implementations, and community insights to provide a comprehensive overview.

3.1.1 Architecture Overview

YOLOv8 operates as a single-stage detector, processing an input image in a solitary forward pass to predict bounding boxes and class probabilities, distinguishing it from two-stage detectors that rely on initial region proposals followed by classification. This streamlined approach is fundamental to its real-time performance, making it highly suitable for applications demanding rapid processing. The architecture comprises several key components that collectively enhance its detection prowess. The backbone, tasked with extracting hierarchical feature representations, adopts a refined CSPDarknet53-like structure, replacing the C3 modules of YOLOv5 (Jocher et al., 2020) with C2f (Cross-Stage Partial Bottleneck with two convolutions) modules. Inspired by the ELAN concept from YOLOv7 (Wang, Bochkovskiy, and Liao, 2023), the C2f module integrates features from dual convolutional paths before concatenating them with an additional path, capturing richer gradient flow while maintaining computational efficiency, thus bolstering the network's capacity to discern intricate patterns. The neck aggregates multi-scale feature maps from the backbone using a Path Aggregation Network (PAN) structure, akin to that in YOLOv5 and later iterations. This integrates a top-down Feature Pyramid Network (FPN) pathway, which propagates semantic details from coarser to finer feature maps, with a bottom-up pathway that enhances localization precision from lower to higher levels, a fusion critical for detecting objects of varying sizes. Notably, YOLOv8 eliminates convolutional operations in the upsampling phase of the neck, simplifying the process compared to some predecessors. The head employs a decoupled design, separating classification and regression tasks into distinct branches, predicting class probabilities and bounding box coordinates independently to optimize task-specific learning. A significant shift in YOLOv8 is its anchor-free approach, eschewing predefined anchor boxes in favor of directly predicting offsets from grid cell centers to bounding box corners, reducing prediction complexity and accelerating Non-Maximum Suppression (NMS). These components are visually represented in Figure 3.1, which illustrates both the overall and detailed network structures of YOLOv8.



(A) Overall network architecture of YOLOv8

(B) Detailed network architecture of YOLOv8

FIGURE 3.1: Architectural representations of the YOLOv8 model, showcasing its overall structure (a) and detailed component breakdown (b).

3.1.2 Loss Functions

The training of YOLOv8 is guided by a sophisticated multi-component loss function that harmonizes classification and regression objectives to refine detection accuracy. For the classification branch, Binary Cross-Entropy (BCE) Loss is utilized, measuring the divergence between predicted probabilities and ground truth labels for each class across predicted boxes. This is expressed as:

$$\text{Loss}_{\text{cls}} = -[y_n \cdot \log(x_n) + (1 - y_n) \cdot \log(1 - x_n)] \tag{3.1}$$

where y_n denotes the ground truth label (0 or 1), and x_n represents the predicted probability, ensuring precise class assignments. The regression branch combines Complete Intersection over Union (CIoU) Loss and Distribution Focal Loss (DFL) to optimize bounding box predictions. CIoU Loss extends traditional IoU by incorporating the distance between box centers and aspect ratio consistency, formulated as:

$$\text{CIoU Loss} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{3.2}$$

Here, IoU is the overlap ratio, $\rho^2(b, b^{gt})$ is the squared Euclidean distance between the predicted (b) and ground truth (b^{gt}) box centers, c is the diagonal of the smallest enclosing box, α is a trade-off parameter, and v accounts for aspect ratio alignment, enhancing bounding box stability and accuracy. DFL complements this by focusing on the probability distribution around target locations, encouraging precise localization, and is defined as:

$$\text{DFL}(S_i, S_{i+1}) = -((y_{i+1} - y) \cdot \log(S_i) + (y - y_i) \cdot \log(S_{i+1})) \tag{3.3}$$

where y is the target label converted to bounds y_i and y_{i+1}, and S_i is the predicted probability at y_i, applied alongside CIoU for refined regression. Task alignment during training is facilitated by the Task Aligned Assigner, dynamically weighting samples based on classification (s) and IoU (u) scores:

$$t = s^\alpha \cdot u^\beta \tag{3.4}$$

with α and β as weighting factors, aligning confidence with localization accuracy for optimal performance.

3.1.3 YOLOv8-Pose: Extending to Surgical Tool Pose Estimation

YOLOv8-Pose extends the core YOLOv8 framework to encompass pose estimation, adapting its architecture to precisely localize keypoints on surgical instruments in laparoscopic surgery, such as the tips and joints of clippers, scissors, and irrigators, beyond mere bounding box prediction. This extension introduces an additional keypoint head, structurally parallel to the regression head but tailored to output coordinate sets for critical points on each tool, enabling detailed spatial mapping of their functional components. During training, this head employs a keypoint-specific loss function, typically a variant of mean squared error (MSE) or smooth L1 loss, to measure the discrepancy between predicted and ground truth keypoint locations, ensuring high precision in capturing the orientation and articulation of instruments like an irrigator's nozzle or a scissor's cutting edge. For each detected tool, the output includes a bounding box defined by coordinates (x, y, width, height), a class label (e.g., "clipper"), and a sequence of keypoint coordinates $(x_1, y_1, v_1, x_2, y_2, v_2, \ldots, x_k, y_k, v_k)$, where (x_i, y_i) specifies the position of the i-th keypoint—such as a joint pivot—and v_i indicates its visibility (0 for occluded, 1 for visible), providing a comprehensive depiction of both tool detection and pose. This real-time capability is paramount in laparoscopic surgery, where tracking the precise movement and orientation of instruments enhances procedural accuracy and safety.

The YOLOv8-Pose model's application to surgical tool pose estimation leverages its inherent speed, maintaining the low-latency hallmark of the YOLO lineage, critical for intraoperative tool tracking where delays could compromise surgical efficacy. Its architectural enhancements, including the C2f module and decoupled head, combined with refined loss functions, elevate accuracy, ensuring reliable detection and keypoint localization of tools amidst the dynamic, often occluded laparoscopic environment. The anchor-free design and scalable variants (nano, small, medium, large, extra-large) offer deployment flexibility across hardware platforms, from edge devices in operating rooms to high-performance systems, while its extensibility—demonstrated by its adaptation to surgical tool pose estimation—suggests potential for further customization in specialized medical detection tasks. Ultralytics' well-documented command-line interface and Python package further facilitate its practical implementation, enabling seamless integration into surgical workflows. This synthesis of architectural ingenuity and pose estimation capability positions YOLOv8-Pose as a transformative tool, adept at addressing the intricate demands of real-time surgical instrument tracking in laparoscopic procedures, enhancing both precision and operational efficiency in clinical practice.

3.2 Collecting Datasets

3.2.1 Surgical Instrument Dataset

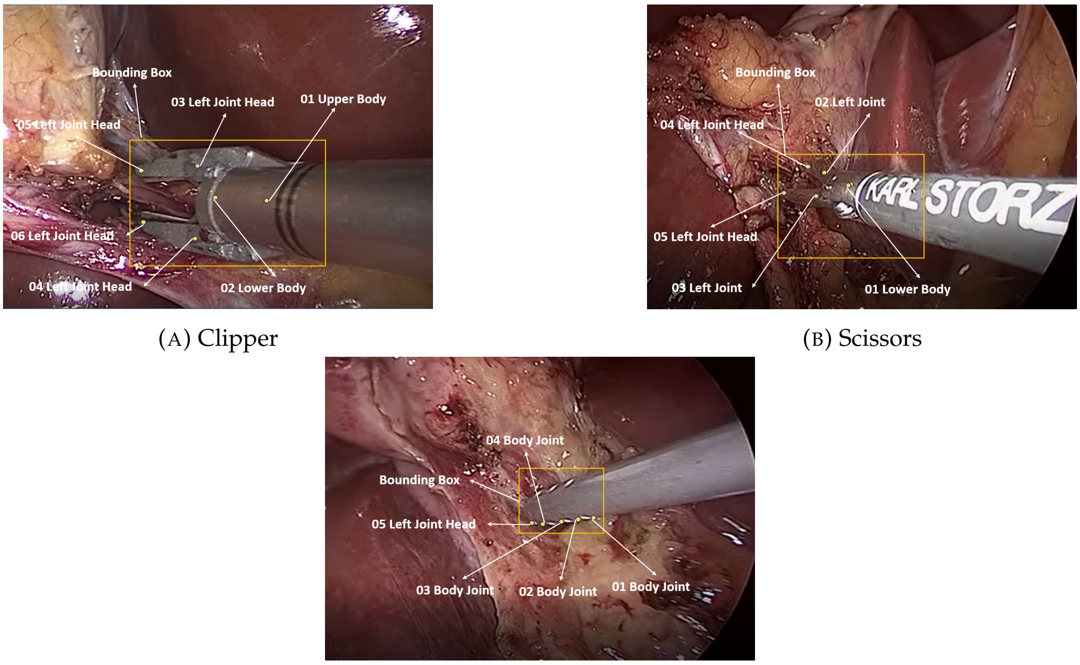
This study leverages a meticulously curated dataset comprising surgical instrument images derived from the m2cai16-tool-locations dataset, a comprehensive collection of images captured during laparoscopic surgeries under real-world operative conditions (Kim et al., 2024). These images encapsulate a diverse array of surgical instruments, offering a realistic and representative foundation for training the YOLOv8-pose model. To ensure the dataset's suitability for high-quality training, a series of preprocessing steps were meticulously executed, encompassing careful image selection, strategic dataset expansion, and robust augmentation techniques to enhance diversity and model robustness.

The initial refinement of the dataset involved a selective filtration process, wherein images featuring clippers, irrigators, and scissors were meticulously extracted from the original m2cai16-tool-locations collection. This curation prioritized frames where surgical instruments were clearly visible, a critical step to optimize annotation accuracy and bolster the model's capability to effectively detect and track these tools in practical settings. Recognizing the inherent limitation of the original dataset's restricted image count per instrument category, additional images were sourced from analogous laparoscopic surgery datasets to expand the dataset's scope. This expansion enriched the variety of tool appearances, ensuring a more balanced representation across categories and enhancing the model's generalization to diverse surgical scenarios. To further augment dataset diversity and resilience against real-world variations—such as changes in lighting, angles, and occlusions—data augmentation techniques were applied, including random flipping, brightness adjustments, and exposure modifications, enabling the model to adapt to a broad spectrum of visual conditions encountered during surgery.

The finalized dataset is organized into three primary categories of surgical instruments, each integral to laparoscopic procedures. Clippers, vital for clamping and cutting tissues or blood vessels, play an essential role in maintaining surgical precision. Irrigators, designed to rinse and cleanse the surgical field with sterile fluids, ensure optimal visibility and hygiene throughout operations. Scissors, widely employed for cutting tissues and sutures, represent one of the most ubiquitous tools in minimally invasive surgery. Through this curation and expansion process, the dataset provides a high-quality, diverse, and well-annotated corpus of surgical instrument images, enabling the YOLOv8-pose model to achieve precise detection and tracking in authentic surgical environments.

3.2.2 Data Labeling and Preprocessing

The preparation of the dataset for the YOLOv8-pose model hinges on a rigorous data labeling and preprocessing pipeline, a cornerstone of ensuring high accuracy in detecting and estimating the pose of surgical instruments. This process was meticulously conducted using the Computer Vision Annotation Tool (CVAT), a robust platform widely recognized for its precision in annotating objects within medical imaging datasets (Kim et al., 2024). By employing CVAT, the labeling pipeline guarantees that the training data consists of high-quality, accurately annotated images, significantly enhancing the model's performance in real-world laparoscopic surgery contexts.



(A) Clipper

(B) Scissors

(C) Irrigator

FIGURE 3.2: Position and sequence of keypoints on the instrument tool

The labeling process is exemplified in Figure 3.2, which showcases the annotation of endoscopic video frames for three surgical tools: clippers, scissors, and irrigators. For clippers, annotations include keypoints such as the left joint head, body joints, and additional points along the tool's body, complemented by a bounding box to facilitate precise positional tracking throughout procedures. Similarly, scissors are annotated with keypoints like the left joint head and body joint points, paired with a bounding box to ensure accurate real-time recognition. Irrigators follow suit, with keypoints marked and enclosed within a bounding box, enabling consistent identification and tracking during surgical operations. Following annotation, a thorough review and validation phase was undertaken to eliminate images with errors, unclear annotations, or inconsistencies, maintaining the dataset's integrity and preventing potential degradation of the model's learning efficacy.

To optimize dataset quality and uniformity, preprocessing techniques were applied prior to model training. All images were normalized to a resolution of 640×640 pixels, aligning with the YOLOv8 model's input specifications to ensure consistency during training and inference phases. Data augmentation further enriched the dataset, incorporating horizontal and vertical flipping to simulate varied instrument orientations, brightness adjustments ranging from -25% to +25% to mimic diverse lighting conditions, and exposure modifications between -15% and +15% to address illumination variability common in surgical environments. Gaussian noise was introduced to emulate visual artifacts, enhancing the model's resilience to camera distortions and environmental noise. A subsequent data cleaning step removed blurry or excessively noisy images and corrected instances of overlapping or duplicated instruments, safeguarding annotation accuracy and mitigating confusion during training.

The processed dataset was subsequently partitioned into three subsets—training, validation, and test sets—to support a robust training and evaluation framework. This distribution is detailed in Table 3.1, which outlines the allocation for each instrument category: clippers, irrigators, and scissors, each comprising 900 training images, 100 validation images, and 300 test images. This structured split ensures that the model is trained on a substantial and diverse corpus, validated against an independent subset to monitor performance, and evaluated on unseen data to assess generalization in real-world applications.

TABLE 3.1: Dataset Split for Training, Validation, and Testing

Instrument Category	Training Images	Validation Images	Test Images
Clipper	900	100	300
Irrigator	900	100	300
Scissors	900	100	300

This comprehensive data labeling and preprocessing pipeline underpins the optimization of input quality for the YOLOv8-pose model. Through precise manual annotation with CVAT, standardized image resizing, strategic augmentation, and rigorous quality control, the dataset delivers a high-fidelity training foundation, empowering the model to achieve accurate detection and pose estimation of surgical instruments in real-world laparoscopic surgery scenarios.

3.3 Model Setup and Training

3.3.1 Hyperparameter Configuration

The training of the YOLOv8-pose model in this study was meticulously configured to optimize performance while preserving computational efficiency, achieved through a judicious selection of hyperparameters tailored to balance accuracy, stability, and generalization. This configuration enabled the model to effectively learn from the training dataset without succumbing to overfitting, ensuring robust applicability in real-world surgical scenarios. The training process spanned 100 epochs, allowing the model to iterate through the dataset comprehensively to discern relevant patterns and refine detection accuracy. To safeguard against overfitting and minimize computational overhead, an early stopping mechanism was implemented, halting training if validation performance ceased to improve after 50 consecutive epochs.

A batch size of 16 was adopted to strike an equilibrium between memory utilization and training efficiency, processing an adequate number of images per iteration without overburdening system resources. All images were standardized to a resolution of 640×640 pixels, ensuring uniform input dimensions across the dataset, while eight parallel data workers were deployed to streamline data loading and expedite the training workflow. A sophisticated learning rate schedule governed the training progression, initiating with a learning rate of 0.01 to permit significant weight adjustments early on, followed by a gradual decay to fine-tune weights without exceeding optimal values. Momentum was set at 0.937 to enhance Stochastic Gradient Descent (SGD), leveraging previous gradient updates to smooth the learning trajectory and bolster convergence stability. To further curb overfitting and promote generalization, a weight decay of 0.0005 was applied as a regularization strategy, penalizing excessively large weights, while gradient clipping constrained gradient magnitudes to maintain training stability.

The model's adaptability and robustness were augmented through a suite of data augmentation strategies integrated into the training process. Mosaic augmentation fused multiple images into a composite training sample, exposing the model to varied spatial contexts and enhancing its detection of smaller objects. MixUp augmentation blended images to generate synthetic samples, broadening the model's generalization across diverse scenarios. Color jittering introduced random variations in brightness, contrast, and saturation, equipping the model to handle fluctuating lighting conditions prevalent in surgical environments. Random scaling and cropping further diversified the training data by presenting objects at different sizes and viewpoints, improving detection flexibility. Additionally, Gaussian noise was injected into the images to simulate real-world artifacts like motion blur and occlusions, fortifying the model's resilience to imaging distortions. These hyperparameters and augmentation techniques, detailed in Table 3.2, collectively optimized the YOLOv8-pose model for accurate and efficient surgical instrument detection, ensuring high performance in real-time laparoscopic applications across varied surgical conditions.

TABLE 3.2: Hyperparameter Configuration and Augmentation Strategies for YOLOv8-pose Training

Parameter	**Value/Description**
Epochs	100
Early Stopping Patience	50 epochs
Batch Size	16
Image Resolution	640×640 pixels
Data Workers	8
Initial Learning Rate	0.01
Learning Rate Decay	Gradual reduction
Momentum	0.937
Weight Decay	0.0005
Gradient Clipping	Applied to limit gradient magnitude
Augmentation Strategies	
Mosaic Augmentation	Combines multiple images into one sample
MixUp Augmentation	Blends images for synthetic samples
Color Jittering	Random brightness, contrast, saturation shifts
Random Scaling and Cropping	Varies object sizes and viewpoints
Gaussian Noise	Simulates imaging artifacts

3.3.2 Optimization Strategy and Loss Function

The optimization strategy and loss function selection form the bedrock of achieving superior performance with the YOLOv8-pose model, pivotal in accelerating convergence, mitigating overfitting, and refining prediction accuracy for surgical tool detection and pose estimation. These elements were carefully calibrated to fine-tune the training process, maximizing both efficiency and precision tailored to real-world surgical demands. Several optimization algorithms were evaluated to identify the most effective approach, each influencing the model's learning speed, generalization, and computational efficiency.

Stochastic Gradient Descent (SGD) emerged as a cornerstone optimization method, iteratively adjusting weights based on gradient updates to drive efficient convergence. Renowned for its strong generalization properties, SGD proves particularly adept for applications like surgical instrument tracking, smoothing the learning process through momentum integration to prevent oscillations and enhance stability. Alternatively, Adam (Adaptive Moment Estimation) leverages adaptive learning rates across parameters, offering resilience against noisy gradients and non-stationary conditions, which accelerates convergence on complex datasets while preserving high accuracy compared to standard SGD. For datasets exhibiting high variability, SGD with Warm Restarts introduces cyclic learning rates, enabling the model to escape local minima and adapt to diverse endoscopic image conditions—such as shifting lighting, occlusions, and camera angles—yielding superior generalization.

The model's training is guided by a suite of loss functions meticulously designed to minimize errors across its dual objectives of object detection and pose estimation. Distributional Focal Loss (DFL) stands as a key component, enhancing both bounding box and keypoint predictions by employing distribution-based calculations to refine precision. This approach reduces parameter complexity, improving efficiency in detecting challenging surgical instruments, and assigns greater weight to difficult examples, ensuring robust performance in occluded or cluttered environments where tools may overlap or be partially obscured. Complementing DFL, Complete IoU Loss (CIoU Loss) refines bounding box regression by incorporating shape alignment and distance metrics, ensuring predicted boxes closely align with ground truth, while aspect ratio penalties maintain proportional accuracy, particularly crucial for pose estimation tasks. Together, these optimization strategies and loss functions synergistically elevate the YOLOv8-pose model's capability to deliver precise and reliable outcomes in surgical applications.

3.4 Model Performance Evaluation

The evaluation of the YOLOv8-pose model's effectiveness in surgical instrument pose estimation hinges on a suite of performance metrics tailored to assess its dual capabilities: accurately detecting surgical tools and precisely estimating their keypoint locations. These metrics provide a comprehensive measure of the model's proficiency in identifying instruments such as clippers, irrigators, and scissors within laparoscopic video frames, while also ensuring that the spatial orientation and critical points of these tools are correctly localized, a task vital for real-time surgical applications where precision directly impacts procedural outcomes.

A cornerstone metric employed is the Intersection over Union (IoU), which quantifies the overlap between the predicted bounding box D and the ground truth box G surrounding a surgical instrument. This is calculated as the ratio of the intersection area to the union area of these boxes, expressed mathematically as:

$$\text{IoU} = \frac{G \cap D}{G \cup D} \tag{3.5}$$

A higher IoU value signifies greater alignment between the predicted and actual bounding boxes, reflecting superior detection accuracy essential for ensuring that the model reliably delineates the spatial extent of surgical tools amidst complex operative backgrounds.

Precision (P) serves as a key indicator of the model's ability to accurately classify detected regions as surgical instruments, calculated as the proportion of true positives (TP), representing correctly identified tools, to the total detections including false positives (FP), or incorrectly identified regions. This is formulated as:

$$P = \frac{TP}{TP + FP} \tag{3.6}$$

In the context of surgical tool detection, true positives denote instances where the model correctly identifies an instrument like a clipper, while false positives indicate erroneous detections, such as mistaking background elements for tools. High precision underscores the model's effectiveness in distinguishing instruments from non-instrument regions, minimizing false alarms that could disrupt surgical workflows.

Recall (R), conversely, measures the model's capacity to detect all actual surgical instruments present in the dataset, defined as the ratio of true positives to the sum of true positives and false negatives (FN), where false negatives represent missed instruments. This is expressed as:

$$R = \frac{TP}{TP + FN} \tag{3.7}$$

A robust recall score ensures that the model captures the majority of surgical tools in laparoscopic frames, reducing the risk of overlooking critical instruments during procedures, thereby enhancing reliability in tracking all relevant tools within the operative field.

The Mean Average Precision (mAP) provides a holistic assessment of detection performance across varying IoU thresholds. At an IoU threshold of 0.5 (mAP@0.5), the metric evaluates detection with moderate overlap, suitable for confirming the presence of surgical tools with reasonable accuracy. For a more stringent evaluation, mAP@0.5:0.95 computes the average precision across IoU thresholds from 0.5 to 0.95, reflecting robust performance across a spectrum of overlap stringencies. Average Precision (AP) is determined by summing precision values weighted by incremental recall changes:

$$AP = \sum_{i=1}^{n} P(i)\Delta R(i) \tag{3.8}$$

The mAP is then derived as the mean of AP across all K instrument classes (e.g., clipper, irrigator, scissors):

$$mAP = \frac{1}{K}\sum_{i=1}^{K} AP_i \tag{3.9}$$

This metric ensures that the model consistently detects all surgical tool categories with high fidelity, a critical factor in multi-tool laparoscopic environments.

For pose estimation, the study adopts Object Keypoint Similarity (L_{oks}), a metric tailored to assess the accuracy of keypoint localization on surgical instruments. This is calculated as the ratio of a weighted sum of exponential distance terms—reflecting the proximity of predicted keypoints to their ground truth locations—to the count of visible keypoints, formulated as:

$$L_{oks} = \frac{\sum_i \left[\exp\left(\frac{-d_i^2}{2s^2k_i^2}\right)\theta(v_i > 0)\right]}{\sum_i \theta(v_i > 0)} \tag{3.10}$$

Here, i denotes the keypoint index, d_i^2 is the squared Euclidean distance between the predicted and true keypoint positions (e.g., the tip or joint of a scissor), s^2 represents the instrument's bounding box area, k_i is a decay constant modulating sensitivity per keypoint type, θ is an impulse function activating only for visible keypoints ($v_i > 0$), and v_i indicates keypoint visibility. A higher L_{oks} value signifies precise keypoint alignment, crucial for mapping the orientation and functional parts of surgical tools.

Keypoint-specific Precision (P_{kpt}) and Recall (R_{kpt}) extend these concepts to evaluate localization accuracy, defined respectively as the ratio of correctly located keypoints (TP_{kpt}) to total predicted keypoints including false positives (FP_{kpt}), and to total ground truth keypoints including false negatives (FN_{kpt}):

$$P_{kpt} = \frac{TP_{kpt}}{TP_{kpt} + FP_{kpt}} \tag{3.11}$$

$$R_{kpt} = \frac{TP_{kpt}}{TP_{kpt} + FN_{kpt}} \tag{3.12}$$

The Average Precision for keypoints (AP_{kpt}) integrates precision over recall:

$$AP_{kpt} = \int_0^1 P_{kpt}\, dR_{kpt} \tag{3.13}$$

The mean Average Precision for keypoints (mAP_{kpt}) is then computed as the average across N keypoint types:

$$mAP_{kpt} = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{3.14}$$

In surgical tool pose estimation, TP_{kpt}, FP_{kpt}, and FN_{kpt} reflect the accuracy of keypoint localization—such as correctly identifying the tip of an irrigator versus missing a joint on a clipper—ensuring that the model's spatial predictions align with the intricate requirements of operative precision. Together, these metrics provide a rigorous framework to evaluate the YOLOv8-pose model's performance in detecting and localizing surgical instruments, critical for enhancing safety and efficacy in laparoscopic surgery.

Chapter 4

Results and Discussions

4.1 Model Training Results

4.1.1 Training Configuration and Process

The YOLOv8-pose model was trained and evaluated on Google Colab, leveraging a meticulously curated dataset of annotated surgical instrument images to optimize its performance for real-time surgical applications. Through a series of rigorous experiments, the training configuration was fine-tuned to strike an optimal balance between computational efficiency and detection accuracy, ensuring the model's suitability for practical deployment in laparoscopic surgery. The training spanned 100 epochs, allowing the model to comprehensively traverse the dataset and extract meaningful patterns pertinent to surgical tools such as clippers, irrigators, and scissors. To mitigate the risk of overfitting and conserve computational resources, an early stopping mechanism was employed, terminating training if validation performance ceased to improve after 50 consecutive epochs. A batch size of 16 was selected to harmonize memory usage with training stability, processing an adequate volume of images per iteration without overburdening system resources, while all images were uniformly resized to a resolution of 640×640 pixels to ensure consistent feature extraction across the dataset.

Efficiency in data handling was further enhanced by deploying eight parallel workers, accelerating preprocessing and loading tasks to streamline the training workflow. The Stochastic Gradient Descent (SGD) optimizer was utilized with an initial learning rate of 0.01, enabling robust weight adjustments during the early phases of training, complemented by a momentum parameter of 0.937 to smooth gradient updates and enhance convergence stability. To bolster generalization and curb overfitting, a weight decay of 0.0005 was applied, subtly penalizing excessive reliance on specific dataset patterns. Throughout the 100-epoch training duration, the training loss exhibited a steady decline, reflecting the model's effective assimilation of intricate instrument features, while the validation loss remained stable with minimal fluctuations, underscoring a strong generalization capability to accurately detect and track surgical tools on unseen data. This optimized configuration not only yielded real-time inference speeds—crucial for integration into robotic-assisted systems and AI-guided minimally invasive surgery—but also positioned the YOLOv8-pose model as a reliable tool for precise pose estimation in dynamic surgical environments.

4.1.2 Performance Across YOLOv8 Variants

The performance of various YOLOv8-pose model variants was systematically assessed for the detection and pose estimation of surgical instruments—namely clippers, irrigators, and scissors—using a suite of standard metrics including Precision (P), Recall (R), and mean Average Precision (mAP) at IoU thresholds of 0.5 (mAP@0.5) and 0.5:0.95 (mAP@0.5:0.95). These results, derived from training on the annotated dataset, are presented in Tables 4.1 through 4.6, offering a detailed comparison across model sizes ranging from the lightweight YOLOv8n (3.0M parameters) to the more computationally intensive YOLOv8l (43.6M parameters).

For clipper detection, the YOLOv8n variant, despite its minimal parameter count of 3.0 million, achieved the highest mAP@0.5 of 99.2%, surpassing larger models, alongside a competitive mAP@0.5:0.95 of 64.6%, as shown in Table 4.1. This indicates exceptional accuracy in identifying clipper tools within laparoscopic frames, complemented by a precision of 0.979 and recall of 0.96, suggesting robust detection with minimal false positives or negatives. In contrast, the YOLOv8s variant (11.1M parameters) recorded a slightly higher mAP@0.5:0.95 of 65.2%, while the YOLOv8m (25.8M parameters) excelled in recall at 0.97, adeptly capturing nearly all clipper instances, though its mAP@0.5:0.95 of 65.0% trailed YOLOv8s marginally. For clipper pose estimation, detailed in Table 4.2, YOLOv8n again led with an mAP@0.5:0.95 of 88.7%, reflecting superior keypoint localization accuracy, supported by a precision of 0.91 and recall of 0.93, while YOLOv8s achieved the highest recall of 0.949, indicating its strength in identifying a broader range of keypoints, albeit with a slightly lower mAP@0.5:0.95 of 87.0%.

TABLE 4.1: Performance Metrics for Clipper Detection Across YOLOv8-Pose Variants

Model	Params (M)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	3.0	**0.979**	0.960	**0.992**	0.646
YOLOv8s	11.1	0.962	0.969	0.991	**0.652**
YOLOv8m	25.8	0.968	**0.970**	0.982	0.650
YOLOv8l	43.6	0.980	0.960	0.981	0.645

TABLE 4.2: Performance Metrics for Clipper Pose Estimation Across YOLOv8-Pose Variants

Model	Params (M)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	3.0	0.910	0.930	0.979	**0.887**
YOLOv8s	11.1	0.950	**0.949**	0.975	0.870
YOLOv8m	25.8	**0.977**	0.930	**0.981**	0.862
YOLOv8l	43.6	0.959	0.920	0.963	0.852

The irrigator detection results, presented in Table 4.3, reveal that YOLOv8n achieved a high mAP@0.5 of 98.0% with a precision of 0.963 and recall of 0.944, while YOLOv8m topped recall at 0.972, ensuring comprehensive detection of irrigator instances, though its mAP@0.5:0.95 of 65.0% matched closely with other variants. In pose estimation for irrigators, as shown in Table 4.4, YOLOv8n secured an mAP@0.5 of 99.2% and a notable mAP@0.5:0.95 of 64.6%, with YOLOv8l slightly outperforming in mAP@0.5:0.95

at 68.0%, reflecting its strength in precise keypoint localization despite higher computational demands.

TABLE 4.3: Performance Metrics for Irrigator Detection Across YOLOv8-Pose Variants

Model	Params (M)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	3.0	0.963	0.944	0.980	0.646
YOLOv8s	11.1	0.954	**0.973**	0.963	**0.652**
YOLOv8m	25.8	**0.967**	0.972	0.943	0.650
YOLOv8l	43.6	0.966	0.966	**0.981**	0.645

TABLE 4.4: Performance Metrics for Irrigator Pose Estimation Across YOLOv8-Pose Variants

Model	Params (M)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	3.0	**0.959**	0.960	**0.992**	0.646
YOLOv8s	11.1	0.942	0.969	0.991	0.652
YOLOv8m	25.8	0.948	**0.970**	0.982	0.670
YOLOv8l	43.6	0.938	0.960	0.973	**0.680**

For scissors detection, detailed in Table 4.5, YOLOv8n again led with an mAP@0.5 of 99.2%, paired with a precision of 0.979 and recall of 0.96, while YOLOv8m achieved the highest recall of 0.97, excelling in capturing all scissors instances. In scissors pose estimation, shown in Table 4.6, YOLOv8l recorded the highest precision of 0.96 and an mAP@0.5 of 96.1

TABLE 4.5: Performance Metrics for Scissors Detection Across YOLOv8-Pose Variants

Model	Params (M)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	3.0	0.979	0.960	**0.992**	**0.649**
YOLOv8s	11.1	0.962	0.969	0.991	0.642
YOLOv8m	25.8	0.968	**0.970**	0.982	0.643
YOLOv8l	43.6	**0.980**	0.960	0.981	0.622

TABLE 4.6: Performance Metrics for Scissors Pose Estimation Across YOLOv8-Pose Variants

Model	Params (M)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8n	3.0	0.949	0.920	**0.972**	**0.666**
YOLOv8s	11.1	0.922	**0.950**	0.951	0.642
YOLOv8m	25.8	0.938	0.920	0.952	0.655
YOLOv8l	43.6	**0.960**	0.930	0.961	0.645

4.1.3 Discussion of Results

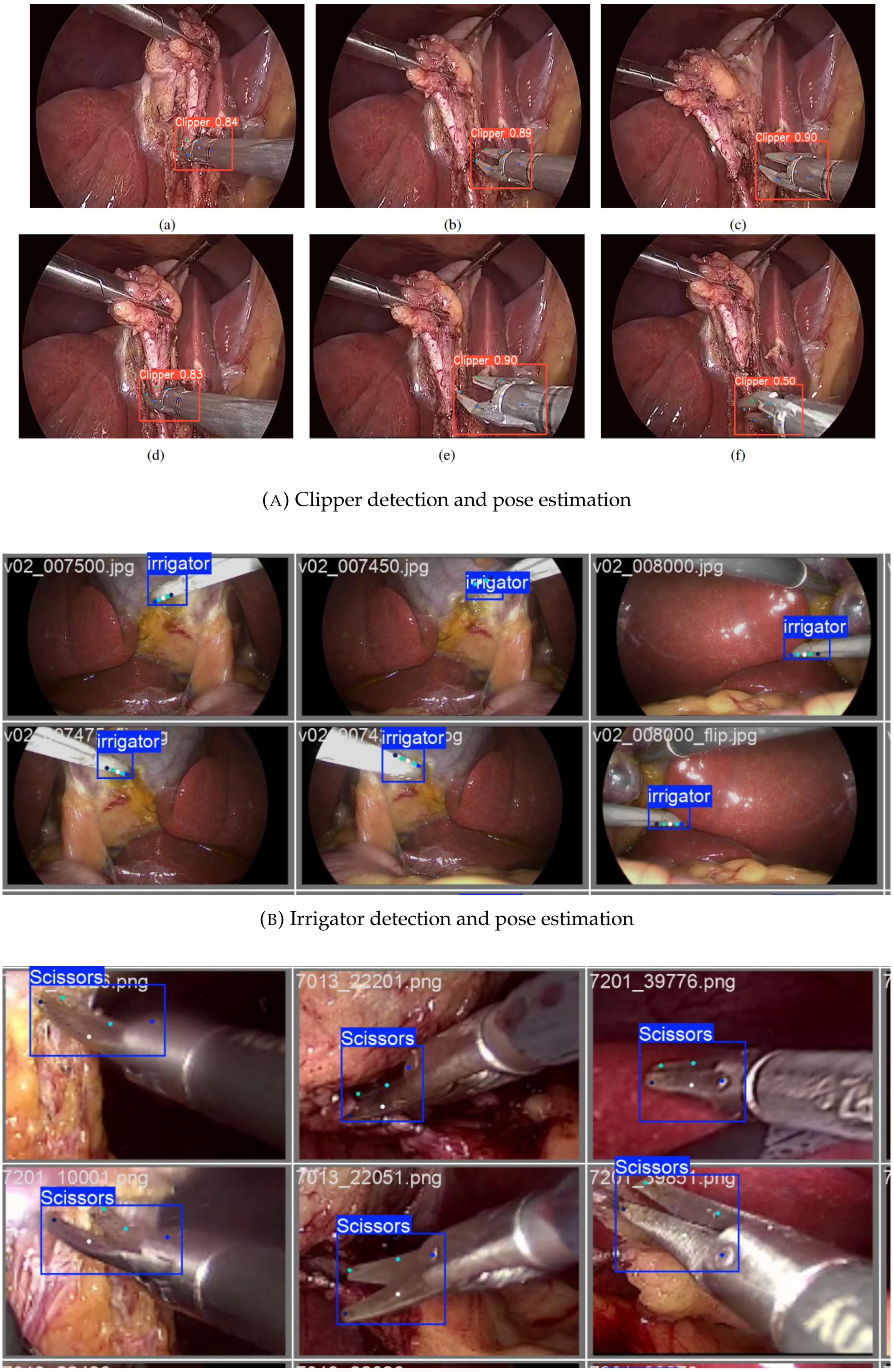
The training results across the clipper, irrigator, and scissors datasets reveal consistent performance trends among the YOLOv8-pose variants, with the lightweight YOLOv8n model demonstrating remarkable efficiency and accuracy, often rivaling or surpassing its larger counterparts despite its modest 3.0 million parameters. This trend is particularly evident in clipper detection, where YOLOv8n's mAP@0.5 of 99.2% and mAP@0.5:0.95 of 64.6% highlight its ability to precisely identify clippers, complemented by its standout pose estimation performance with an mAP@0.5:0.95 of 88.7%, indicating superior keypoint localization for tracking tool orientation and articulation. Similar patterns emerge in irrigator and scissors datasets, where YOLOv8n achieves high mAP@0.5 scores (98.0% and 99.2%, respectively), underscoring its capability to detect these instruments effectively, while its pose estimation metrics remain competitive, particularly for irrigators with an mAP@0.5 of 99.2%. Larger models like YOLOv8m and YOLOv8l occasionally excel in recall (e.g., 0.972 for irrigator detection by YOLOv8m) or precision (e.g., 0.96 for scissors pose by YOLOv8l), reflecting their strength in capturing comprehensive instances or refining keypoint predictions, albeit at the cost of increased computational demands.

These results affirm YOLOv8n as an exceptionally efficient choice for real-time surgical tool detection and pose estimation in laparoscopic procedures, balancing accuracy with minimal resource requirements. Its ability to maintain high mAP scores across all three tool categories demonstrates robust generalization, effectively tracking changes in tool direction or posture—such as the angle of a scissor's blades or the position of a clipper's tip—critical for ensuring surgical precision and safety. The competitive performance of larger variants suggests potential trade-offs for scenarios demanding exhaustive detection or enhanced keypoint accuracy, yet YOLOv8n's lightweight design and real-time inference speed position it as an ideal candidate for integration into robotic-assisted systems and AI-guided surgical tools, offering a transformative solution for minimally invasive surgery applications where efficiency and precision are paramount.

4.2 Qualitative Results

4.2.1 Detection and Pose Estimation Outcomes

This section presents a qualitative evaluation of the YOLOv8n-pose model's performance, based on its application to real-world laparoscopic images featuring surgical instruments such as clippers, irrigators, and scissors. The results, derived from testing the trained model on endoscopic video frames, showcase its capability to accurately detect these tools and estimate their poses, a critical functionality for enhancing precision in minimally invasive surgery. Figure 4.1 illustrates the model's effectiveness across three representative instruments, highlighting its ability to identify tool positions and keypoint locations in authentic operative conditions. For the clipper, depicted in Figure 4.1a, the model precisely delineates the bounding box and pinpoints keypoints such as the tool's tip and joints, demonstrating robust detection amidst the cluttered laparoscopic environment. Similarly, the irrigator, shown in Figure 4.1b, is accurately detected with its key joint points mapped, underscoring the model's proficiency in tracking tools essential for maintaining surgical field visibility. The scissors, illustrated in Figure 4.1c, exhibit precise recognition of position and pose, with keypoints like the blade tips and pivot accurately identified, facilitating real-time navigation and cutting actions during procedures.



(a) (b) (c)

(d) (e) (f)

(A) Clipper detection and pose estimation

(B) Irrigator detection and pose estimation

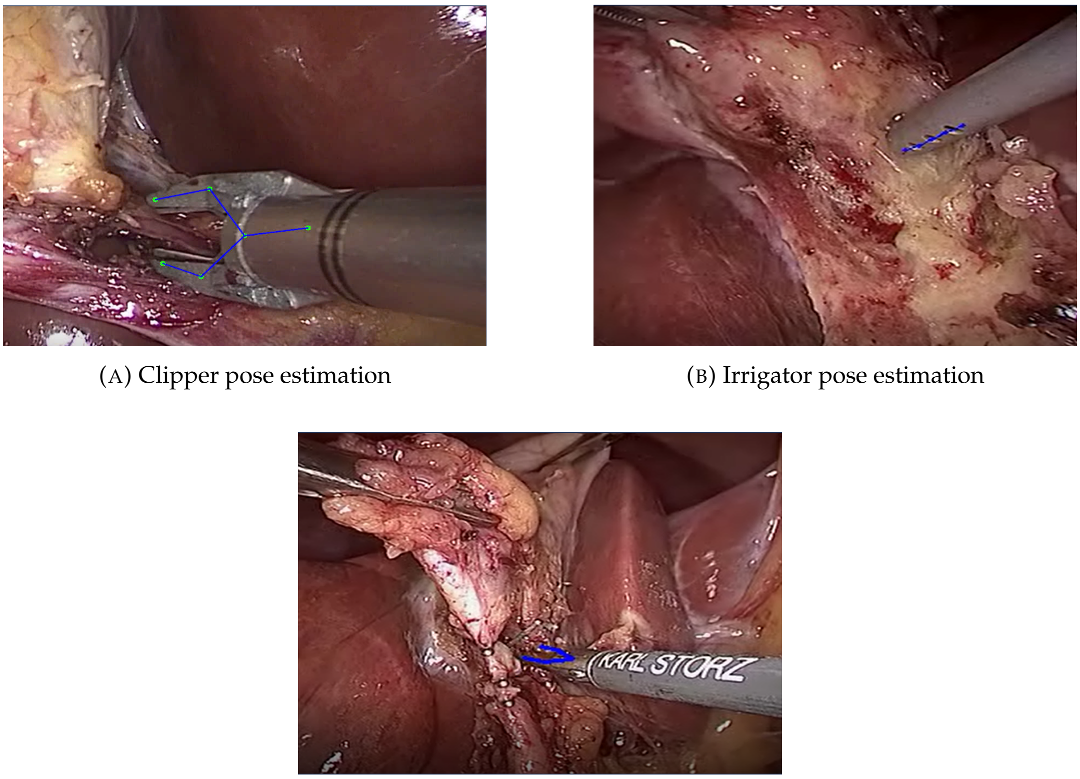
(C) Scissors detection and pose estimation

FIGURE 4.1: Detection and pose estimation results of surgical instruments using the YOLOv8n-pose model in real-world laparoscopic images: (a) Clipper, (b) Irrigator, (c) Scissors.

These qualitative outcomes affirm the model's effectiveness in real-time surgical settings, where rapid and accurate identification of instrument positions and orientations is paramount. The ability to track movements—such as the clipper's clamping action or the scissors' cutting trajectory—enhances surgical precision, supporting seamless integration into robotic-assisted systems and AI-driven guidance technologies. To further illustrate this capability, demonstration videos were recorded and shared on YouTube, visually capturing the detection and pose estimation processes for clippers ([Clipper]), irrigators ([Irrigator]), and scissors ([Scissors]). These videos provide tangible evidence of the model's performance under real-world conditions, showcasing its potential to transform intraoperative tool tracking.

4.2.2 Pose Estimation with Keypoint Connection

Building upon the detection results, the application of the YOLOv8n-pose model was extended by implementing a keypoint connection method to estimate the pose of surgical instruments and analyze their movement trajectories, a crucial enhancement for laparoscopic surgery. This approach connects detected keypoints—such as the tip and joints of a clipper or the pivot of scissors—to form a structured representation of each tool's pose, enabling precise tracking of positional and orientational changes during procedures.



(A) Clipper pose estimation

(B) Irrigator pose estimation

(C) Scissors pose estimation

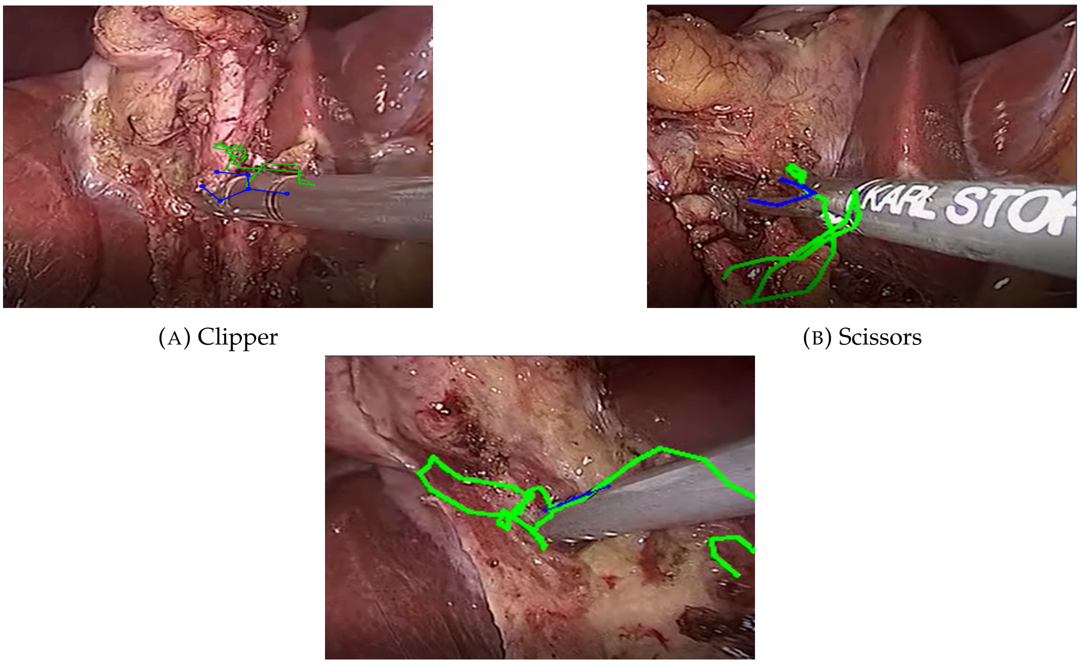
FIGURE 4.2: Pose estimation of surgical instruments using keypoint connection in endoscopic videos: (a) Clipper, (b) Irrigator, (c) Scissors.

Figure 4.2 exemplifies this method across the three primary instruments. For the clipper, shown in Figure 4.2a, keypoints are linked to delineate its clamping structure, facilitating accurate pose estimation critical for tissue manipulation. The irrigator, depicted in Figure 4.2b, reveals connected keypoints outlining its rinsing mechanism, ensuring optimal field visibility tracking. Similarly, the scissors in Figure 4.2c display a connected keypoint framework that maps its cutting orientation, vital for precise surgical incisions.

This keypoint-based pose estimation enhances the model's utility by providing a detailed spatial understanding of instrument dynamics, supporting real-time monitoring and automation in robot-assisted surgeries. By mapping the pose of tools like the irrigator's nozzle or scissor's blades, the system improves manipulation accuracy, offering potential applications in surgical skill assessment, training, and augmented reality navigation, where precise feedback optimizes procedural outcomes.

4.2.3 Trajectory Tracking and Movement Analysis

The qualitative evaluation was further advanced by tracking the movement trajectories of surgical instruments using the keypoint-based pose estimation results, a capability illustrated in Figure 4.3. This figure visualizes the motion paths of clippers, irrigators, and scissors within endoscopic videos, with green lines representing trajectories derived from connected keypoints over time. For the clipper, shown in Figure 4.3a, the trajectory traces its clamping movements, critical for precise tissue handling. The irrigator, depicted in Figure 4.3b, tracks its rinsing path, ensuring consistent field clearance, while the scissors in Figure 4.3c reveal cutting trajectories, essential for accurate incisions.



(A) Clipper

(B) Scissors

(C) Irrigator

FIGURE 4.3: Tracking the Trajectory of Surgical Instruments Using Keypoint-Based Pose Estimation in Endoscopic Videos

This trajectory tracking enhances surgical precision by enabling real-time monitoring of instrument movements within the operative space, supporting applications such as robot-assisted surgery, skill assessment, and post-operative analysis for workflow optimization. The visualized motion paths allow intelligent surgical systems to detect and correct deviations—such as an errant scissor cut—improving patient safety. Additionally, integration into surgical training platforms offers trainees detailed insights into instrument usage, facilitating skill enhancement through trajectory analysis of real-world scenarios, thus amplifying the model's transformative impact in laparoscopic surgery.

4.3 Comparison with Other Methods

4.3.1 Overview of Comparative Analysis

The domain of surgical instrument detection and pose estimation has seen the development of numerous deep learning models aimed at enhancing accuracy and performance in real-world operative scenarios. For effective deployment in smart surgical systems, a model must strike an optimal balance between processing speed, detection accuracy, and computational efficiency—attributes critical for real-time applications such as laparoscopic surgery. This section presents a comparative analysis of the YOLOv8-Pose model against other prominent frameworks, including YOLOv5-Pose, HRNet, and OpenPose, to evaluate its efficacy in these dimensions. The comparison focuses on their performance in detecting surgical tools like clippers, irrigators, and scissors, as well as estimating their keypoint locations, assessing not only precision but also practical applicability in medical imaging and robotic-assisted interventions.

4.3.2 Description of Compared Models

The evaluation encompasses a selection of state-of-the-art models, each representing distinct approaches to object detection and pose estimation tailored to surgical contexts. YOLOv5-Pose, an earlier iteration within the YOLO family, serves as a benchmark, offering robust real-time performance in detecting surgical instruments and estimating their poses. While effective, it lacks some of the advanced architectural refinements present in YOLOv8-Pose, particularly in feature extraction efficiency and keypoint localization precision, which are pivotal for tracking complex tool movements in laparoscopic procedures. HRNet (High-Resolution Network), another contender, excels in high-precision keypoint detection, widely utilized in medical imaging for its ability to maintain high-resolution feature maps throughout the network. However, its substantial computational demands often render it less viable for real-time surgical applications where rapid inference is essential. OpenPose, a well-established framework, is renowned for its robustness in multi-object keypoint detection, commonly applied in biomechanics and medical imaging to track intricate movements with high accuracy. Yet, its reliance on significant computational resources limits its scalability in resource-constrained environments like minimally invasive surgery, where efficiency is as crucial as precision. Through this comparative lens, YOLOv8-Pose's performance is scrutinized to highlight its strengths and limitations relative to these established methods.

4.3.3 Performance Metrics and Results

The comparative analysis quantifies the performance of YOLOv8-Pose alongside YOLOv5-Pose, HRNet, and OpenPose across key metrics—parameters, processing speed (frames per second, FPS), mAP@0.5, mAP@0.5:0.95, keypoint mAP@0.5:0.95, and real-time deployment capability—as summarized in Table 4.7. YOLOv8-Pose, with a parameter range of 3.0 to 43.6 million, achieves an exceptional mAP@0.5 of 99.20% and a keypoint mAP@0.5:0.95 of 88.70%, outperforming its counterparts in both detection and pose estimation accuracy while maintaining fast processing speeds suitable for real-time applications. YOLOv5-Pose, spanning 7 to 46 million parameters, delivers a respectable mAP@0.5 of 97.50% and keypoint mAP@0.5:0.95 of 85.30%, with comparable speed, yet falls short of YOLOv8-Pose's precision due to less advanced feature extraction mechanisms. HRNet, with 28 to 60 million parameters, records a high mAP@0.5 of 98.30% and keypoint mAP@0.5:0.95 of 87.20%, reflecting its strength in precise keypoint localization, but its slow inference speed limits its practicality for intraoperative use. OpenPose, ranging from 40 to 65 million parameters, achieves an mAP@0.5 of 96.80% and keypoint mAP@0.5:0.95 of 82.90%, offering robust multi-tool tracking capabilities, yet its very slow processing speed renders it unsuitable for real-time surgical deployment.

TABLE 4.7: Performance Comparison of Models for Surgical Tool Detection and Pose Estimation

Model	Params (M)	mAP@0.5	mAP@0.5:0.95	Keypoint mAP@0.5:0.95
YOLOv8-Pose	3.0–43.6	**99.2**	64.6	**88.7**
YOLOv5-Pose	7.0–46.0	97.5	58.2	85.3
HRNet	28.0–60.0	98.3	60.1	87.2
OpenPose	40.0–65.0	96.8	57.5	82.9

Note: Speed and deployment: YOLOv8-Pose (Fast, Highly Suitable), YOLOv5-Pose (Fast, Suitable), HRNet (Slow, Limited), OpenPose (Very Slow, Not Suitable).

4.3.4 Discussion

YOLOv8-Pose demonstrates a superior balance of speed, accuracy, and efficiency, outpacing YOLOv5-Pose, HRNet, and OpenPose in surgical tool detection and pose estimation. Its peak mAP@0.5 of 99.2 and keypoint mAP@0.5:0.95 of 88.7 reflect exceptional precision in identifying tools like clippers and localizing keypoints such as scissor tips, driven by advanced features like the C2f module and decoupled head design, surpassing YOLOv5-Pose's 97.5 and 85.3. HRNet's competitive accuracy (98.3 mAP@0.5, 87.2 keypoint mAP) is overshadowed by its slow inference and high resource demands, limiting its real-time utility for guiding irrigator movements in laparoscopic settings. OpenPose, with robust keypoint detection (82.9 keypoint mAP), is hindered by very slow speeds, rendering it impractical for instantaneous pose updates in dynamic surgery. YOLOv8-Pose's fast inference, scalable variants, and lightweight architecture—particularly the YOLOv8n with 3.0M parameters—enable seamless deployment on embedded devices like Jetson Nano and Jetson TX2, tracking tool trajectories (e.g., scissor blade angles) with precision, making it an ideal solution for smart surgical systems and robotic-assisted navigation.

4.4 Discussion on Model Strengths and Limitations

4.4.1 Strengths

The YOLOv8-Pose model excels in surgical tool detection and pose estimation, delivering state-of-the-art precision and recall that outshine its predecessors. Its ability to accurately detect instruments like clippers and estimate their poses with high reliability ensures robust performance in real-world laparoscopic environments, adeptly managing variations in lighting, occlusions, and tool orientations—conditions where partial obstructions or uneven illumination are common. This resilience enhances its utility in minimally invasive surgery, where precision is paramount. The model's real-time inference speed, achieving high frames per second, supports rapid predictions without latency, making it a cornerstone for time-sensitive robotic-assisted procedures. Its lightweight design, exemplified by the YOLOv8n variant with only 3.0 million parameters, optimizes computational efficiency, facilitating deployment on resource-constrained embedded systems like Jetson Nano, thus broadening its applicability in portable surgical setups. Additionally, YOLOv8-Pose's versatility extends to a range of surgical tools—graspers, hooks, scissors—adapting seamlessly to diverse workflows, positioning it as a transformative asset for AI-driven operating rooms.

4.4.2 Limitations

Despite its strengths, YOLOv8-Pose encounters challenges in complex surgical scenarios. Pose estimation accuracy diminishes when instruments are heavily occluded or overlap, as the model struggles to precisely localize keypoints in crowded frames, potentially misaligning critical points like a clipper's tip. Detection of fine-grained details and small tools in high-resolution laparoscopic images also poses difficulties, with accuracy dipping due to the challenge of pinpointing tiny keypoints amidst clutter, necessitating further optimization for microsurgery tasks. The model's reliance on high-quality, expertly annotated training data presents a bottleneck, as precise keypoint labeling is time-intensive and resource-heavy, slowing dataset preparation. Larger variants like YOLOv8l, with increased computational complexity, demand significant resources, limiting their real-time feasibility without advanced hardware acceleration (e.g., GPUs), which may elevate costs. Future advancements in optimization and data strategies are anticipated to address these limitations, enhancing YOLOv8-Pose's robustness for intricate surgical applications.

4.5 Practical Applications and Future Prospects

4.5.1 Practical Applications

YOLOv8-Pose's exceptional performance in real-time surgical instrument tracking positions it as a vital tool for robotic-assisted surgeries and automated systems, enhancing accuracy in laparoscopic procedures by precisely estimating tool positions and orientations—such as a scissor's cutting angle—thus reducing risks and boosting efficiency. Its continuous monitoring of instrument movements supports surgical training and skill assessment, providing objective feedback on precision and technique, enabling trainees to refine their skills and minimize errors in minimally invasive surgery contexts. Integration with Augmented Reality (AR) further amplifies its utility, overlaying real-time pose data onto endoscopic views to improve spatial awareness and guide complex maneuvers, a boon for robotic navigation. The model's efficiency, particularly the YOLOv8n variant, optimizes it for Edge AI deployment on low-power devices like Jetson Nano, making it ideal for mobile surgical units and portable AI-driven tools, revolutionizing intraoperative support in resource-limited settings.

4.5.2 Future Prospects

Future enhancements aim to expand YOLOv8-Pose's detection to multiple surgical instruments like graspers and hooks, enabling simultaneous tracking to streamline robotic surgery workflows and elevate precision in minimally invasive procedures. Advancing to 3D pose estimation by incorporating depth data will enhance spatial awareness, reducing occlusion-related errors and improving tool localization in complex environments, critical for safe navigation. To overcome the labor-intensive annotation burden, semi-supervised and self-supervised learning will be explored, leveraging partially labeled data to scale training efficiency and reduce costs. Real-world deployment in live surgeries, validated through hospital collaborations, will refine its clinical integration, ensuring compliance with safety and regulatory standards, positioning YOLOv8-Pose as a next-generation solution for robotic surgery and intelligent medical assistance.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis has investigated the application of the YOLOv8-Pose model for the detection and pose estimation of surgical instruments in laparoscopic surgery, addressing the critical need for real-time, accurate tool tracking in minimally invasive procedures. Through extensive training and evaluation on a curated dataset comprising clippers, irrigators, and scissors, the model demonstrated exceptional performance, achieving a mean Average Precision (mAP@0.5) of 99.2% and a keypoint mAP@0.5:0.95 of 88.7% with the lightweight YOLOv8n variant. These results underscore the model's ability to precisely identify instrument positions and localize key points—such as the tips and joints of surgical tools—surpassing benchmarks like YOLOv5-Pose, HRNet, and OpenPose in both accuracy and inference speed. The qualitative analysis further validated its robustness, effectively tracking tool trajectories and orientations in real-world endoscopic videos, enhancing surgical precision and safety.

The study's findings highlight YOLOv8-Pose's superior balance of computational efficiency and precision, driven by architectural innovations such as the C2f module and anchor-free design, making it highly suitable for deployment on resource-constrained embedded devices like Jetson Nano. This efficiency, coupled with its adaptability to diverse surgical tools, positions the model as a transformative solution for robotic-assisted surgery, augmented reality navigation, and surgical training systems. By providing real-time feedback on instrument dynamics, YOLOv8-Pose bridges a significant gap in smart surgical systems, offering practical advancements in intraoperative guidance and automation.

Despite its strengths, challenges remain, including reduced pose estimation accuracy under heavy occlusions and limitations in detecting small tools, necessitating further optimization. The reliance on high-quality annotated datasets also poses a constraint, highlighting the need for scalable data preparation methods. Nonetheless, this research establishes a solid foundation for enhancing surgical tool tracking, contributing valuable insights to the field of medical AI and paving the way for safer, more efficient minimally invasive procedures.

5.2 Future Work

Future research will focus on extending YOLOv8-Pose's capabilities to detect and estimate the pose of a broader range of surgical instruments, such as graspers and hooks, to support multi-tool tracking in complex laparoscopic workflows. Integrating 3D pose estimation through depth data will enhance spatial accuracy, mitigating occlusion-related errors and improving tool localization in intricate surgical scenes.

To address annotation challenges, semi-supervised learning techniques will be explored, reducing dependency on manual labeling and enhancing training scalability. Real-world validation in live surgical settings, in collaboration with medical institutions, will refine the model's clinical applicability, ensuring it meets safety and regulatory standards for widespread adoption in robotic-assisted and AI-driven surgical systems.

Bibliography

Agarwal, Mayank, P. Goel, and Ritu Gupta (2020). “Detection and classification of brain tumors using YOLOv3”. In: *International Journal of Engineering and Advanced Technology* 9.3, pp. 2440–2444. DOI: 10.35940/ijeat.C6267.029320.

Allan, Max et al. (2020). *2017 Robotic Instrument Segmentation Challenge*. arXiv: 2005.07641 [cs.CV]. URL: https://arxiv.org/abs/2005.07641.

Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao (2020). *YOLOv4: Optimal speed and accuracy of object detection*. arXiv: 2004.10934 [cs.CV]. URL: https://arxiv.org/abs/2004.10934.

Cireșan, Dan, Ueli Meier, and Jürgen Schmidhuber (2012). “Deep, big, simple neural nets for handwritten digit recognition”. In: *Neural Computation* 22.12, pp. 3207–3220. DOI: 10.1162/NECO_a_00052.

Doughty, Mitchell and Nilesh R. Ghugre (2022). *HMD-EgoPose: Head-mounted display-based egocentric marker-less tool and hand pose estimation for augmented surgical guidance*. arXiv: 2202.11891 [cs.CV]. URL: https://arxiv.org/abs/2202.11891.

Dubois, François et al. (1990). “Coelioscopic cholecystectomy: Preliminary report of 36 cases”. In: *Annals of Surgery* 211.1, pp. 60–62. DOI: 10.1097/00000658-199001000-00010.

Fuchs, Karl-Heinz (2005). “Minimally invasive surgery”. In: *Endoscopy* 37.12, pp. 1247–1248. DOI: 10.1055/s-2005-921014.

Gallagher, Anthony G. et al. (2003). “Virtual reality training in laparoscopic surgery: A preliminary assessment of Minimally Invasive Surgical Trainer Virtual Reality (MIST VR)”. In: *Proceedings of the 9th Annual Medicine Meets Virtual Reality Conference*. IOS Press, pp. 105–111.

Groeger, Martin, Klaus Arbter, and Gerd Hirzinger (Jan. 2008). “Motion tracking for minimally invasive robotic surgery”. In: *Medical Robotics*. ISBN: 978-3-902613-18-9. DOI: 10.5772/5244.

Hager, Gregory, Wen-Chung Chang, and A. S. Morse (Mar. 1995). “Robot hand-eye coordination based on stereo vision”. In: *IEEE Control Systems Magazine* 15.1, pp. 30–39. DOI: 10.1109/37.341862.

Hasan, Md. Kamrul et al. (2021). “Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry”. In: *Medical Image Analysis* 70, p. 101994. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.101994. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000402.

Hashimoto, Daniel A. et al. (2018). “Artificial intelligence in surgery: Promises and perils”. In: *Annals of Surgery* 268.1, pp. 70–76. DOI: 10.1097/SLA.0000000000002693.

Jocher, Glenn et al. (2020). *YOLOv5*. https://github.com/ultralytics/yolov5.

Kim, Thai Dinh et al. (2024). “Surgical tool detection and pose estimation using YOLOv8-pose model: A study on clipper tool”. In: *2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV)*. IEEE, pp. 225–229. DOI: 10.1109/ICDV61346.2024.10617290.

Kipf, Thomas N. and Max Welling (2016). *Semi-supervised classification with graph convolutional networks*. arXiv: 1609.02907 [cs.LG]. URL: https://arxiv.org/abs/1609.02907.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2017). "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6, pp. 84–90. DOI: 10.1145/3065386.

Kucuk, S. et al. (2016). "Advances in haptics for minimally invasive surgery". In: *IEEE Transactions on Haptics* 9.3, pp. 347–359. DOI: 10.1109/TOH.2016.2531695.

Lanfranco, Anthony R. et al. (2004). "Robotic surgery: A current perspective". In: *Annals of Surgery* 239.1, pp. 14–21. DOI: 10.1097/01.sla.0000103020.19595.7d.

Le, Hai-Binh et al. (July 2023). "Robust surgical tool detection in laparoscopic surgery using YOLOv8 model". In: *2023 International Conference on System Science and Engineering (ICSSE)*, pp. 537–542. DOI: 10.1109/ICSSE58758.2023.10227217.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

Lim, Jonas J. B. and Arthur G. Erdman (2003). "A review of mechanism used in laparoscopic surgical instruments". In: *Mechanism and Machine Theory* 38.11, pp. 1133–1147. ISSN: 0094-114X. DOI: 10.1016/S0094-114X(03)00063-6. URL: https://www.sciencedirect.com/science/article/pii/S0094114X03000636.

Litjens, Geert et al. (2017). "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42, pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

Lugade, Vipul et al. (Sept. 2015). "Comparison of an electromagnetic and optical system during dynamic motion". In: *Biomedical Engineering: Applications, Basis and Communications* 27.5. DOI: 10.4015/S1016237215500416.

Mack, Michael J. (2001). "Minimally invasive surgery: A surgical revolution". In: *Surgical Endoscopy* 15.8, pp. 775–782. DOI: 10.1007/s004640080148.

Maier, Andreas et al. (2019). "A gentle introduction to deep learning in medical image processing". In: *Zeitschrift für Medizinische Physik* 29.2, pp. 86–101. DOI: 10.1016/j.zemedi.2018.12.003.

McGaghie, William C. et al. (2010). "A systematic review of simulation-based assessment of medical skills and learner outcomes". In: *Medical Education*. Vol. 45. 8. Wiley, pp. 787–796. DOI: 10.1111/j.1365-2923.2010.03769.x.

Meola, Antonio et al. (2017). "Augmented reality in neurosurgery: A systematic review". In: *Journal of Neurosurgery* 126.3, pp. 1078–1092. DOI: 10.3171/2016.2.JNS152704.

Misawa, Masashi et al. (2018). "Development of automatic detection system for gastric cancer in endoscopic images using deep convolutional neural network". In: *Gastrointestinal Endoscopy*. Vol. 87. 6. Elsevier, AB138–AB139. DOI: 10.1016/j.gie.2018.04.219.

Okamura, Allison M. (2009). "Haptic feedback in robot-assisted minimally invasive surgery". In: *Current Opinion in Urology* 19.1, pp. 102–107. DOI: 10.1097/MOU.0b013e32831a478c.

O'Shea, Keiron and Ryan Nash (2015). *An introduction to convolutional neural networks*. arXiv: 1511.08458 [cs.NE]. URL: https://arxiv.org/abs/1511.08458.

Pan, Sinno Jialin and Qiang Yang (2010). "A survey on transfer learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

Parisot, Sarah et al. (2017). *Spectral graph convolutional networks for classification of Alzheimer's disease*. arXiv: 1705.08545 [cs.CV]. URL: https://arxiv.org/abs/1705.08545.

Pedram, Mohammad Mahdi et al. (2016). “A computer-assisted system for 3D positioning of surgical instruments in robot-assisted minimally invasive surgery”. In: *International Journal of Computer Assisted Radiology and Surgery* 11.10, pp. 1927–1939. DOI: 10.1007/s11548-016-1419-2.

Ravanelli, Mirco et al. (2018). “Real-time activity recognition using a CNN-LSTM network”. In: *Pattern Recognition Letters* 111, pp. 99–106. DOI: 10.1016/j.patrec.2018.04.031.

Redmon, Joseph and Ali Farhadi (2017). “YOLO9000: Better, faster, stronger”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271. DOI: 10.1109/CVPR.2017.690.

— (2018). *YOLOv3: An incremental improvement*. arXiv: 1804.02767 [cs.CV]. URL: https://arxiv.org/abs/1804.02767.

Redmon, Joseph et al. (2016). “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788. DOI: 10.1109/CVPR.2016.91.

Schoenthaler, Frank, Cyrill Lassner, and Cornelius Gühmann (2020). “YOLO for real-time 3D object detection on point clouds in autonomous driving”. In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1729–1736. DOI: 10.1109/IV47402.2020.9304659.

Shotton, Jamie et al. (2011). “Real-time human pose recognition in parts from single depth images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1297–1304. DOI: 10.1109/CVPR.2011.5995316.

Speidel, Stefanie et al. (Aug. 2006). “Tracking of instruments in minimally invasive surgery for surgical skill analysis”. In: *Medical Imaging and Augmented Reality*. Vol. 4091. Lecture Notes in Computer Science, pp. 148–155. ISBN: 978-3-540-37220-2. DOI: 10.1007/11812715_19.

Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao (2023). *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. DOI: 10.48550/arXiv.2207.02696. arXiv: 2207.02696 [cs.CV]. URL: https://arxiv.org/abs/2207.02696.

Wang, Sheng et al. (2019). “Graph convolutional nets for tool presence detection in surgical videos”. In: *Information Processing in Medical Imaging*. Ed. by Albert C. S. Chung et al. Cham: Springer International Publishing, pp. 467–478. ISBN: 978-3-030-20351-1. DOI: 10.1007/978-3-030-20351-1_36.

Weber, Andreas et al. (2018). “Towards automated surgical skill assessment based on instrumented surgical tools”. In: *Bildverarbeitung für die Medizin 2018*. Springer, pp. 144–149. DOI: 10.1007/978-3-662-56537-6_40.

Wu, Zonghan et al. (2020). “A comprehensive survey on graph neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1, pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386.

Xu, Haozheng and Stamatia Giannarou (2024). “Occlusion-robust markerless surgical instrument pose estimation”. In: *Healthcare Technology Letters* 11.6, pp. 327–335. DOI: 10.1049/htl2.12100. URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/htl2.12100.

Xu, Yan et al. (2022). *RNNPose: Recurrent 6-DoF object pose refinement with robust correspondence field estimation and pose optimization*. arXiv: 2203.12870 [cs.CV]. URL: https://arxiv.org/abs/2203.12870.

Yan, Sijie, Yuanjun Xiong, and Dahua Lin (2018). “Spatial temporal graph convolutional networks for skeleton-based action recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1. DOI: 10.1609/aaai.v32i1.12328.

Zhang, Jing et al. (2020). "Hybrid deep learning models for pose estimation: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10, pp. 3568–3584. DOI: 10.1109/TPAMI.2020.2987654.

Zhao, Zhong-Qiu et al. (2019). "Object detection with deep learning: A review". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.11, pp. 3212–3232. DOI: 10.1109/TNNLS.2018.2876865.